Design and analysis of MAC protocols in CDMA wireless networks

Zhu, Qichao

2006

Zhu, Q. (2006). Design and analysis of MAC protocols in CDMA wireless networks. Master's thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/4082>

<https://doi.org/10.32657/10356/4082>

Nanyang Technological University

Downloaded on 26 Apr 2023 19:25:00 SGT

9416162

Design and Analysis of MAC Protocols in CDMA Wireless Networks

Zhu Qichao

School of Electrical and Electronic Engineering

A thesis submission to Nanyang Technological University
in fulfillment of the requirement for the degree of
Master of Engineering

2006

TK
5105.55
Z63
2006

Statement of Originality

I hereby certify the content of this thesis is the result of work done by me and has not been submitted for a higher degree to any other University or Institution.

Date

Zhu Qichao

Acknowledgements

I am greatly indebted to my supervisor, Prof. Ma Maode, for his knowledge, instruction, dedication and exemplary academic standards. He always provides thoughtful guidance that contributes a lot to this thesis. I am deeply impressed by his rigorous attitude to do research that will have effect on my future research and work.

I am deeply grateful to my best friends, Wang Xinliang and Zhou Yin Zhu, for their concern, help and encouragement. The consistent support from my parents is deeply appreciated.

Comfortable studying environment provided by Nanyang Technological University is sincerely appreciated as it supports the completion of my study. In particular, I am thankful to the technicians Mr. Lim Puay Chye and Ms. Lim Yoke Lan in Communication Lab IV for their technical support in the use of computers, software, and other equipments.

Special thanks go to one of my old friends. It is who that helps to extricate me out of depravation and freshens me to face the difficulties in life.

Table of Contents

Acknowledgements i

Table of Contents..... ii

Summary v

List of Figures..... vii

List of Tables xi

Symbols xii

Abbreviations..... xiv

Chapter 1 Introduction 1

 1.1 Background Information..... 1

 1.2 Motivation 4

 1.3 Contribution..... 10

 1.3.1 Providing Integrated Service in CDMA Wireless Networks..... 10

 1.3.2 Providing Differentiated Service in CDMA Wireless Networks 11

 1.3.3 Thesis Organization 12

Chapter 2 Literature Review 14

 2.1 MAC Protocols for Wireless CDMA Networks..... 14

 2.1.1 Random Access MAC Protocols 15

 2.1.2 Random Reservation Protocols 16

 2.1.3 Demand Assignment MAC Protocols 21

2.2	A Near-Optimum MAC Protocol Based on DQRAP for a CDMA Mobile Communication System.....	31
2.2.1	System Background.....	31
2.2.2	Protocol Description.....	34
Chapter 3	Providing Integrated Service in CDMA Wireless Networks.....	38
3.1	Motivation	38
3.2	System Model and Service	40
3.3	Scheduling Issue and Algorithms.....	42
3.3.1	Collision Resolution Scheme.....	43
3.3.2	Data Transmission Queue Scheduling.....	44
3.4	Performance Evaluation	49
3.4.1	Experimental Design	50
3.4.2	Experiment Results.....	56
Chapter 4	Providing Differentiated Services in CDMA Wireless Networks.....	77
4.1	Introduction	77
4.2	Proportional Delay Differentiation.....	79
4.2.1	PDD model	79
4.2.2	Performance of the PDD Model	81
4.2.3	Features of the PDD Model	82
4.3	Hierarchical Scheduling for Integrated Traffic in CDMA Wireless Networks..	85
4.3.1	System Model and Service	85
4.3.2	Scheduling Issues and Algorithm	87
4.4	Admission Control Algorithm.....	104
4.4.1	Admission Control in Wireless CDMA Networks	104
4.4.2	Proposed Admission Control Algorithm	106
4.5	Performance Evaluation	118
4.5.1	Experiment for Scheduling Algorithms.....	118
4.5.2	Experiment for Proposed Admission Control	138

Chapter 5 Conclusions and Recommendations 148

 5.1 Conclusions 148

 5.1.1 Providing Integrated Service in CDMA Networks..... 148

 5.1.2 Providing Differentiated Service in CDMA Networks 149

 5.2 Recommendations 150

 5.2.1 Integrated Service 150

 5.2.2 Differentiated Service..... 151

Author’s Publication..... 152

 Published 152

 In Submission 152

Bibliography 153

APPENDIX 158

 A Brief Introduction to SSS..... 158

Summary

The design of Medium Access Control (MAC) protocols plays a very important role in the design of wireless networks. In this thesis, we study two issues, integrated service and differentiated service, with aim to provide QoS service to mixed traffic composed of real-time and non-real-time messages. First, a novel MAC protocol named as distributed queuing random access with assignment protocol (DQRA+AP/CDMA) has been proposed to provide real-time service to the time-constrained applications in a CDMA wireless network. DQRA+AP/CDMA has been designed based on distributed queuing random access protocol for CDMA wireless network (DQRAP/CDMA) with consideration of providing real-time service to an integrated traffic. The major contributions of the proposed protocol are to: 1) combine the random access scheme with the fixed assignment method to reduce the collisions among the transmission requests when they access control subslots in order to support real-time service; 2) introduce a time-based priority scheme to enhance real-time service; 3) propose a simple scheduling scheme to support real-time and non-real-time traffic; 4) further introduce a cost-based scheduling algorithm to enhance the performance of the integrated service with certain fairness. The results obtained from the simulation experiments reveal that the proposed protocol and scheduling algorithms can provide real-time services with good performance while improving the performance of the service to the non-real-time traffic.

The second issue is to provide proportional delay differentiated (PDD) service between real-time and non-real-time traffic classes. We have introduced the original proportional

average delay (PAD) scheduler as the inter-class scheduler to meet the PDD constraints when they are feasible. The PAD scheduler has been modified to provide the average delay differentiation between the real-time and non-real-time classes. Furthermore, one intra-class scheduler has been designed for each of both classes according to their particular requirements. Finally, based on the proposed hierarchical scheduler composed of inter- and intra-class schedulers, we have proposed an admission control algorithm to control the number of the traffic flows accepted to the network so that the QoS requirements of the real-time traffic can be met while the network utilization can be kept as high as possible. The results obtained from the simulation experiments indicate that the inter- and intra-class schedulers can work effectively side by side. The inter-class scheduler can provide the delay differentiated service to the real-time and non-real-time traffic under the feasible conditions. The intra-class schedulers can improve the absolute performance of the network in terms of average message delay and tardy rate as expected. In addition, the admission control algorithm is able to provide guaranteed QoS service to time-constrained messages, and at the same time, keep the network utilization at a high level.

List of Figures

Figure 1.1 OSI Layers Relevant for the Air Interface	3
Figure 2.1 Model of DQRAP/CDMA.	37
Figure 3.1 System Model of Distributed Queues	41
Figure 3.2 Flow Chart of Simulation Program for DQRA+AP/CDMA/MUF+FIFO	53
Figure 3.3 Flow Chart of Simulation Program for DQRA+AP/CDMA/MUFC	54
Figure 3.4 Flow Chart of Simulation Program for DQRA+AP/CDMA/CBPA	55
Figure 3.5 Average Real-time Message Loss Rate vs. Total Traffic Load	57
Figure 3.6 Average Real-time Message Delay vs. Total Traffic Load.....	58
Figure 3.7 Average Real-time Message Loss Rate vs. Total Traffic Load	59
Figure 3.8 Average Message Delay vs. Total Traffic Load	60
Figure 3.9 Average non-real-time Message Delay vs. Total Traffic Load.....	60
Figure 3.10 Average Message Loss Rate vs. Traffic Load when Real-time Traffic Percentage is 20%.....	61
Figure 3.11 Average Message Loss Rate vs. Traffic Load when Real-time Traffic Percentage is 40%.....	62
Figure 3.12 Average Message Loss Rate vs. Traffic Load when Real-time Traffic Percentage is 60%.....	62
Figure 3.13 Average Message Loss Rate vs. Traffic Load when Real-time Traffic Percentage is 80%.....	63
Figure 3.14 Average Message Loss Rate vs. Traffic Load when Real-time Traffic Percentage is 100%.....	63
Figure 3.15 Average Message Delay vs. Traffic Load when Real-time Traffic Percentage is 0%	65
Figure 3.16 Average Message Delay vs. Traffic Load when Real-time Traffic Percentage is 20%	65
Figure 3.17 Average Message Delay vs. Traffic Load when Real-time Traffic Percentage is 40%	65

Figure 3.18 Average Message Delay vs. Traffic Load when Real-time Traffic Percentage is 60% 66

Figure 3.19 Average Message Delay vs. Traffic Load when Real-time Traffic Percentage is 80% 66

Figure 3.20 Average Message Delay vs. Traffic Load when Real-time Traffic Percentage is 100% 67

Figure 3.21 Average Real-time Message Loss Rate vs. Traffic Load with 60% Real-time Traffic 68

Figure 3.22 Average Non-real-time Message Delay vs. Traffic Load with 60% Real-time Traffic 69

Figure 3.23 Average Real-time Message Loss Rate vs. Traffic Load with 80% Real-time Traffic 70

Figure 3.24 Average Non-real-time Message Delay vs. Traffic Load with 80% Real-time Traffic 71

Figure 3.25 Average Real-time Message Loss Rate vs. Traffic Load with 20% Real-time Traffic 72

Figure 3.26 Average Non-real-time Message Delay vs. Traffic Load with 20% Real-time Traffic 73

Figure 3.27 Average Real-time Message Delay vs. Traffic Load when Real-time traffic percentage is 20% 74

Figure 3.28 Average Real-time Message Delay vs. Traffic Load when Real-time traffic percentage is 40% 74

Figure 3.29 Average Real-time Message Delay vs. Traffic Load when Real-time traffic percentage is 60% 75

Figure 3.30 Average Real-time Message Delay vs. Traffic Load when Real-time traffic percentage is 80% 75

Figure 3.31 Average Real-time Message Delay vs. Traffic Load when Real-time traffic percentage is 100% 76

Figure 4.1 Structure of Scheduling Algorithm Combined of Inter- and Intra-Class Schedulers 88

Figure 4.2 A Queue with a Waiting Buffer and a Transmission Queue 99

Figure 4.3 System Model of the Proposed Protocol 100

Figure 4.3A Procedures of the Propose Admission Control Algorithm..... 114

Figure 4.3B Procedures of the Propose Admission Control Algorithm for New Requests
..... 117

Figure 4.4 Average Message Delay vs. Total Message Arrival Rate When Real-time
Traffic Percentage = 20% 119

Figure 4.5 Average Message Delay vs. Total Message Arrival Rate When Real-time
Traffic Percentage = 50% 120

Figure 4.6 Average Message Delay vs. Total Message Arrival Rate When Real-time
Traffic Percentage= 80% 120

Figure 4.7 Real-time and Non-real-time Message Delay vs. Real-time Traffic Percentage
and Aggregate Arrival Rate 121

Figure 4.8 Average Message Delay vs. Total Message Arrival Rate When Delay Ratio =4
..... 124

Figure 4.9 Average Message Delay vs. Total Message Arrival Rate When Delay Ratio = 8
..... 124

Figure 4.10 Average Message Delay vs. Total Message Arrival Rate When Real-time
Traffic Percentage = 10% 125

Figure 4.11 Average Message Delay vs. Total Message Arrival Rate When Real-time
Traffic Percentage = 30% 125

Figure 4.12 Average Message Delay vs. Total Message Arrival Rate When Real-time
Traffic Percentage = 50% 126

Figure 4.13 Average Message Delay vs. Total Message Arrival Rate When Real-time
Traffic Percentage = 70% 126

Figure 4.14 Average Message Delay vs. Total Message Arrival Rate When Real-time
Traffic Percentage = 90% 127

Figure 4.15 Real-time Message Average Tardy Rate vs. Real-time Message Arrival Rate
..... 129

Figure 4.16 Real-time Message Average Delay vs. Real-time Message Arrival Rate.... 129

Figure 4.17 Non-real-time Message Average Delay vs. Non-Real-time Message Arrival
Rate 130

Figure 4.18 Average Message Delay vs. Aggregate Arrival Rate When Real-time Traffic
Percentage = 20% 131

Figure 4.19 Average Message Delay vs. Aggregate Arrival Rate When Real-time Traffic Percentage = 50% 132

Figure 4.20 Average Message Delay vs. Aggregate Arrival Rate When Real-time Traffic Percentage = 80% 132

Figure 4.21 Real-time and Non-real-time Message Delay vs. Real-time Traffic Percentage and Aggregate Arrival Rate 133

Figure 4.22 Real-time Message Tardy Rate vs. Aggregate Arrival Rate When Real-time Traffic Percentage = 20% 135

Figure 4.23 Real-time Message Tardy Rate vs. Aggregate Arrival Rate When Real-time Traffic Percentage = 50% 136

Figure 4.24 Real-time Message Tardy Rate vs. Aggregate Arrival Rate When Real-time Traffic Percentage= 80% 136

Figure 4.25 Real-time Tardy Rate vs. Real-time Traffic Percentage and Aggregate Arrival Rate 137

Figure 4.26 Real-time Message Delay vs. Real-time Traffic Percentage and Aggregate Arrival Rate 138

List of Tables

Table 4.1 Average Messages Delay When the Aggregate Arrival Rate =15 Packets/Slot 121

Table 4.2 Average Messages Delay When the Aggregate Arrival Rate =15 Packets/Slot 128

Table 4.3 Average Messages Delay When the Aggregate Arrival Rate =15 Packets/Slot 134

Table 4.4 Parameters for the Traffic Flows from the Mobile Nodes. 142

Table 4.5 Parameters for the Traffic Flows from the Mobile Nodes. 144

Table 4.6 Parameters for the Traffic Flows from the Mobile Nodes. 146

Symbols

λ_i	the average arrival rate of class i
λ_{RT}	the average aggregate arrival rate of real-time class
λ_{NRT}	the average aggregate arrival rate of non-real-time class
d^m	the time constraint, or laxity, of message m
d_i	mean laxity of real-time traffic flow i
$\bar{d}_i$	the average delay of class i
$\tilde{d}_i$	the normalized average delay of class i
$\tilde{d}_i^n$	the normalized laxity of message n in class i
$\bar{D}_{RT}$	average message delay for whole real-time class
D^k	the deadline of message k
$D_i(t)$	the sequence of messages arriving at subqueue i during every t
P_i	the number of all departed messages $D_i(t)$
S_i	the sum of delays of all departed messages in $D_i(t)$
$P^i(t)$	the priority of i th message in queue
P_{RT}^i	the priority of i th message in real-time class
P_{NRT}^i	the priority of i th message in non-real-time class
w^m	the experienced waiting time of message m in the system
$w_i(t)$	the head waiting time of class i

$\tilde{w}_i(t)$	the normalized head waiting time of class i
a^k	the arriving time of message k
M^n	the transmission time of message n
$\tilde{M}^n$	the normalized transmission time of message n
$\tilde{M}_i^n$	the normalized transmission time of message n in class i
l_i	the average length of traffic flow i
l^m	the length of message m
$R_{td}(t)$	the tardy rate of real-time queue at time t
$N_{UN}(t)$	the number of the unsuccessful transmitted real-time messages till now
$N_{NRT}(t)$	then number of total real-time messages served till now
N_i	average number of messages in the buffer for traffic flow i
Tr_{total}	the total traffic load of the network
Tr_{RT}	the traffic load of real-time messages
Tr_{NRT}	the traffic load of non-real-time messages
Per_{RT}	percentage of real-time traffic in the network
δ_i	the delay differentiation parameter of class i
μ	the total service capacity of the network

Abbreviations

ABR	Available Bit Rate
AP	Access Point
ARIB	the Japanese Association of Radio Industries and Business
ATM	Asynchronous Transfer Mode
BER	Bit Error Rate
CAC	Call Admission Control
CBPA	Cost-based Priority Assignment
CBR	Constant Bit Rate
CDMA	Code Division Multiple Access
CSMA/CD	Carrier Sense Multiple Access with Collision Detection
DDP	Delay Differentiation Parameter
DiffServ	Differentiated Service
DQRAP/CDMA	Distributed Queuing Random Access Protocol for CDMA
DQRA+AP/CDMA	Distributed Queuing Random Access with Assignment Protocol for CDMA
DQRUAM	Distributed Queue Request Update Multiple Access
DS	Differentiated Service
DSCP	Differentiated Service Code Point
DSMA	Direct Sequence Multiple Access
DTR	Date Transmission Rules
EDF	Earliest Deadline First

ETSI	European Telecommunication Standard Institute
FCFS	First-come-first-serve
FDD	Frequency Division Duplexing
FHMA	Frequency Hopped Multiple Access
FIFO	First-in-first-out
GPS	General Processor Sharing
GSEBF	Gated Smallest Estimated Bandwidth First
HOL	Head-of-the-line
IID	Identical Distribution
IMT-2000	International Mobile Telecommunications-2000
IS	Integrated Service
ISMA	Inhibit Sense Multiple Access
ITU	International Telecommunication Union
LIFO	Last-in-first-out
MAC	Medium Access Control
MS	Mobile Station
MSC	Mobile Switching Center
MUF	Most Urgent First
MUFC	Most Urgent First with Comparison
MUF-FIFO	MUF with First-in-first-out
NRT	Non-real-time
PAD	Proportional Average Delay
PDD	Proportional Delay Differentiation
PHB	Per-hop Behavior

PN	Pseudo-noise
PRMA	Packet Reservation Multiple Access
QDR	Queuing Discipline Rules
QoS	Quality of Service
RACH	Random Access Channel
RT	Real-time
RTR	Request Transmission Rules
SAFD	Smallest Arrival First Drop
SEBF	Smallest Estimated Bandwidth First
SEBFD	Smallest Estimated Bandwidth First Drop
SIR	Signal-to-interference
SNR	Signal-noise-ratio
SJF	Shortest Job First
SPTF	Shortest Processing Time First
SSMA	Spread Spectrum Multiple Access
TDD	Time Division Duplexing
TDP	Time-dependent Priority
UBR	Unspecified Bit Rate
UMTS	Universal Mobile Telecommunication System
VBR	Variable Bit Rate
WCDMA	Wideband CDMA

Chapter 1 Introduction

1.1 Background Information

Spread Spectrum Multiple Access (SSMA) uses the signal that has a transmission bandwidth with several orders of magnitude greater than the minimum required radio frequency bandwidth. A pseudo-noise (PN) sequence converts a narrowband signal to a wideband noise-like signal before transmission. SSMA also provides immunity to multipath interference with robust multiple access capability. SSMA is not very bandwidth efficient when used by a single user. However, since many users can share the same spread spectrum bandwidth without interfering with one another, spread spectrum systems become bandwidth efficient in a multiple user environment. It is exactly this feature that has attracted wireless system designers. There are two major types of spread spectrum multiple access techniques: *frequency hopped multiple access* (FH) and *direct sequence multiple access* (DS). Direct sequence multiple access is also called code division multiple access (CDMA).

In CDMA systems, the narrowband message signal is multiplied by a very large bandwidth signal called as a spreading signal. The spreading signal is a pseudo-noise code sequence with a chip rate that is orders of magnitudes greater than the data rate of the message. All users in a CDMA system use the same carrier frequency and may transmit simultaneously. Each user has its own pseudorandom codeword that is approximately orthogonal to all other codewords. The receiver performs a time correlation operation to detect only the specific desired codeword. All other codewords

appear as noise due to decorrelation. For detection of the message signal, the receiver needs to know the codeword used by the transmitter. Each user operates independently with no knowledge of the other users.

The features of CDMA including the following:

- Many users of a CDMA system can share the same frequency. Either time division duplexing (TDD) or frequency division duplexing (FDD) may be used.
- Unlike TDMA or FDMA, CDMA has a soft capacity limit. Increasing the number of users in a CDMA system raises the noise floor in a linear manner. Thus, there is no absolute limit on the number of users in CDMA. Rather, the system performance gradually degrades for all users as the number of users increase, and gets better when the number of users decreases.
- Multipath fading could be substantially reduced because the signal is spread over a large spectrum. If the spread spectrum bandwidth is greater than the coherence bandwidth of the channel, the inherent frequency diversity will mitigate the effects of small-scale fading.
- Channel data rates are very high in CDMA systems. Consequently, the symbol (chip) duration is very short and usually much less than the channel delay spread. Since PN sequences have low autocorrelation, multipath that is delayed by more than a chip will appear as noise. A RAKE receiver can be used to improve reception by collecting time delayed versions of the required signal.
- Since CDMA uses co-channel cells, it can use macroscopic spatial diversity to provide soft handoff. Soft handoff is performed by the mobile switching center (MSC), which can simultaneously monitor a particular user from two or more base

stations. The MSC could choose the best version of the signal at any time without switching frequencies.

- Self-jamming is a problem in CDMA system. Self-jamming arises from the fact that the spreading sequences of different users are not exactly orthogonal. Hence in the spreading of a particular PN code, non-zero contributions to the receiver decision statistic for a desired users arise from the transmissions of other users in the system.

It appears that the problems of multiple access can readily be split into the sub-problems of:

- Providing a *basic multiple access scheme* such as CDMA; and
- Choosing a suitable set of rules, a so-called *multiple access protocol* on top of that.

The basic scheme would be associated with the first or lowest OSI layer, the physical layer, while the multiple access protocol is commonly situated at the lower sub-layer of the second layer, the so called medium access protocol (MAC) layer (Figure 1.1).

Figure 1.1 OSI Layers Relevant for the Air Interface

1.2 Motivation

In cellular communication systems, a multi-access or shared channel and consequently a multiple access protocol are required at least:

- To allow mobile users to register in the system (e.g., when switching their handsets on);
- For mobile users to send occasional location update messages. This enables the network to track users and to limit sending pages (i.e., notifications of incoming calls) in cells of the appropriate location area rather than all cells of the network;
- To allow users (or rather terminals) to place a request for resources to make a call. This could be either a user initiated (i.e., *mobile originated* call), or as a response to a page, i.e., a *mobile terminated* call. Upon reception of such requests, the base station will attempt to reserve the required resources and notify the user of the resources to use and any potentially temporal restrictions regarding the use of these resources.

The number of codes per cell available for users may be limited, in which case codes cannot be provided on a per-user-basis, but only on a per-call-basis. Even when plenty of codes are available, the base station cannot decode user signals without having a rough idea on what codes it is expected to use. Therefore, an access mechanism is required for mobiles to request codes. The logical channel used for this purpose is typically some type of random access channel, on which a requesting user may pick one of a limited number of codes known to the base station.

One of the driving forces of the next generation of wireless communication and computing networks is the promised high-speed multimedia services. Third-generation

systems, such as the International Mobile Telecommunication System (IMT-2000) network (formerly known as the Future Public Land Mobile Telecommunication System, FPLMTS) and the Universal Mobile Telecommunication System (UMTS), promise to provide multimedia services to mobile and fixed users via wireless access to the global telecommunications infrastructure. The IMT-2000 is a universal, multifunction, globally compatible digital mobile radio system that plans to integrate all traffic types and all wireless radio systems under a common set of formats. The standards for IMT-2000 are being developed by the International Telecommunication Union (ITU) standards organization for global communications. UMTS is a similar global wireless solution, and is being standardized by the European Telecommunications Standards Institute (ETSI). Among the requirements for third-generation systems is the ability to support multimedia traffic. The various classes of traffic can be distinguished as suggested by the Asynchronous Transfer Mode (ATM) Forum:

- Constant bit rate (CBR) traffic (digital voice and video)
- Real-time variable bit rate (rt-VBR) traffic (compressed voice and video)
- Non-real-time variable bit (nrt-VBR) traffic (data)
- Available bit rate (ABR) traffic (non-time-critical data)
- Unspecified bit rate (UBR) traffic (file transfer, system backup, email)

Each of these classes has its unique traffic characteristics and quality-of-service (QoS) requirements. Due to the heterogeneous and bursty nature of multimedia traffic flows, the traditional voice-based medium access control (MAC) protocols provide best effort service. They cannot meet the emergent demands to support the multiservice with different QoS requirements in one network. Thus flexible MAC protocols that can

efficiently accommodate multimedia traffic are required for third-generation systems. To serve the requirements of multimedia traffic, a MAC protocol must:

- Provide simultaneous support for the wide variety of traffic types mentioned above
- Support traffic that requires delay and jitter bounds
- Assign bandwidth resources in an efficient manner (e.g., on demand)
- Support both fair and prioritized access to resources

One important issue of MAC protocols is the message scheduling or queuing discipline. The policies that schedule the order of transmissions for messages with time constraints will have a great impact on the efficiency and performance of MAC protocols for future wireless networks. However, the design of a message scheduler involves balancing a number of conflicting requirements. Common criteria for messages scheduler design include maximization of throughput, QoS provisioning, scheduling according to a predefined priority structure, and less complexity. The decision to choose which message in the queue is to be served next may be based on any or all of the following:

- some measure related to the relative arrival times for those customers in the queue;
- some measure (exact value, estimate, pdf) of the service time required or the service so far received;
- or some function of group membership.

Examples of queuing disciplines that depend only upon arrival time are First-in-first-out (FIFO), Last-in-first-out (LIFO), and random order of service. Discrimination on the basis of service time only may take the following forms: Shortest-job-first (SJF), Longest-job-

first (LJF), similar rules based on averages, and so on. Order of service based on an externally imposed priority class structure may take many forms as, for example, the Head-of-the-line (HOL) system.

Random access protocols have been widely used in the past for wireless networks. Message reservation multiple access (PRMA) is a well-known TDMA based protocol proposed for voice and data traffic [13]. For CDMA-based systems, the protocol proposed in [2] can be regarded as a pure random access protocol, while the protocols presented in [3], [4], [5] and [6] are in close relationship with PRMA [13]. They are the combination of random access and reservation-based resource sharing methods. The mobile terminals in the protocols presented in [7-12] can obtain the transmission resources after their requests compete successfully.

Distributed queuing random access protocol (DQRAP) is a protocol that uses three control minislots and resolves collisions based on a tree-typed algorithm [15], [16]. It was initially designed for a TDMA environment. Its performance can approach that of an ideal M/D/1 queue, reaching maximum stable throughputs close to one, and maintaining its stability for traffic loads up to channel capacity. An improvement of DQRAP can be found in [17], [18], and [19], where a near-optimum protocol based on DQRAP for CDMA wireless network has been proposed, named as DQRAP/CDMA. This protocol can behave like a random access protocol when the traffic load is light and switch automatically to a reservation protocol when the traffic load increases. DQRAP/CDMA can achieve a very good performance in supporting data traffic. However, it has not considered supporting real-time service for an integrated traffic in the CDMA wireless

networks. Integrated traffic is composed of real-time streams, in which messages have time constraints, and non-real-time streams without time constraints. The most important aspect of the time-constrained application is that a message generated at a source station must be received correctly at its destination station within a given amount of time. This period is regarded as a message deadline. If the delay of a message in the network exceeds its time constraint, the message is considered as late. A scheduling algorithm for a MAC protocol that serves real-time messages is to schedule the transmission of the messages to meet their time constraints as much as possible. However, the service that benefits the real-time traffic may cause unfairness to the non-real-time streams of the integrated traffic in the wireless networks. The challenge to the design of scheduling algorithms is how to serve the real-time messages without violation of their deadlines while providing transmission service to the non-real-time messages with reasonable performance in order to reach some extends of fairness.

Another issue in the wireless access network is differentiated services. The Differentiated Services (DS) architecture is proposed as a response to the scalability problems in the Integrated Services (IS) concept. DS architecture reduces the state of information stored in the network compared to the IS architecture, by providing QoS to limited number of classes.

Two approaches for service differentiation have been defined: *absolute differentiated services* and *relative differentiated services*. The first relies on admission control and resource reservation mechanisms in order to provide guarantees and statistical assurances for performance measures, such as minimum service rate and end-to-end delay. The

approach of relative differentiated services groups network traffic into N classes of service which are ordered, such that Class I is better (or at least no worse) than Class $(i-1)$ for $1 < i \leq N$, in terms of local (per-hop) metrics for the queuing delays and message losses. The eight *Class Selector PHBs*, standardized by the IETF, follow the relative service differentiation model. The network assures that higher classes will offer better QoS than lower classes. It is up to the applications and users to select the class that best meets their requirements, cost, and policy constraints. From this point of view, the relative service differentiation model follows the architectural principle of *end-system adaptation*, since it provides the class selection as an additional dimension in the end-system adaptation space.

The basic premise for a relative DS architecture that it has to be:

- *Predictable*, in the sense that the differentiation should be consistent (i.e., higher classes are better, or at least no worse) independent of the class load variations
- *Controllable*, meaning that the network operators should be able to adjust the relative QoS difference among classes based on their criteria.

As a target for predictable and controllable relative differentiation, we consider the *proportional differentiation model*. According to this model, the basic performance measures for message forwarding locally at each hop are ratioed proportionally to certain *class differentiation parameters* that the network operators choose. Even though there is no wide consensus on the most appropriate performance measures for message forwarding, it is generally agreed that a better network service means lower queuing delays and lower likelihood of message losses. We apply the proportional model in the

case of queuing delay differentiation, which is according to the important performance, laxity (delay), of the real-time traffic in the transmission.

1.3 Contribution

1.3.1 Providing Integrated Service in CDMA Wireless Networks

We propose a novel MAC protocol, named as Distributed Queuing Random Access with Assignment Protocol (DQRA+AP/CDMA) based on the DQRAP/CDMA protocol to provide real-time service to the time-constrained applications in a CDMA wireless network. The first contribution of the proposed protocol is to combine the random access scheme in the original DQRAP/CDMA protocol with the fixed assignment method to reduce the collisions among the transmission requests when they access control minislots in order to support real-time service. Furthermore, a time-based priority scheme, named as Most Urgent First (MUF), is introduced to schedule the waiting messages in the data transmission queue in order to enhance real-time service for the time-constrained traffic. In short term, it allocates higher priorities to the more urgent messages so that it can reduce the average message loss rate because of missing the deadline before transmission. In long term, the queuing discipline schedules the messages in an earlier arrival to transmit before those in a later arrival. This long-term First-In-First-Out (FIFO) order decreases the average message delay due to MUF method that causes longer delays to the messages with larger laxities. Then we extend the proposed algorithm to support integrated traffic composed of real-time and non-real-time traffic. It transmits the messages (real-time and non-real-time) in an earlier arrival before those in a later arrival while in the same batch the real-time messages have absolutely higher priorities than the non-real-time messages. Based on the above proposed protocol two more complicated

scheduling algorithms, Most Urgent First with Comparison (MUFC) and Cost-based Priority Assignment (CBPA), are presented for more reasonable balance on the performance of the network under an integrated traffic. In MUFC the real-time messages are not allocated absolutely higher priorities than the non-real-time messages in the same batch. It inserts at most one non-real-time message to transmit between the successive deliveries of the two real-time messages if this insertion does not lead the second real-time message to miss the deadline. The CBPA algorithm schedules the integrated traffic in a more delicate way that it considers the deadline and processing time of the messages jointly to determine the priorities and schedules the real-time and non-real-time messages in a more conform way.

1.3.2 Providing Differentiated Service in CDMA Wireless Networks

In the design of the scheduler for differentiated service in the wireless network, we divide the proposed scheduler in two parts. One is the inter-class scheduler to support the differentiated service between the real-time and non-real-time classes. The other is intra-class scheduler to manage the messages in the same class with respect to the particular requirements. The inter-class and intra-class schedulers work in the different levels. We choose the proportional average delays (PAD) scheduler to provide service differentiation among the classes because of its flexibility. Since PAD is unaware of the waiting times of backlogged messages, however, it occasionally allows higher classes to experience much larger delays than their long-term average delays, or the queuing delays of lower classes. Therefore, we introduce a window measurement method to the original PAD scheduler so that the modified attempts to equalize the short-term normalized average delays in the last K departures but not all the previous departures until present. Thus, it in a great degree reduces the abnormal behavior that original PAD exhibits in the short timescales and

reflects the burstiness in the message network traffic variation. The intra-class schedulers of real-time and non-real-time traffic are different. In the design of the intra-class scheduler for real-time class, we propose a MUF scheduling algorithm combined with Shortest Processing Time First (SPTF) method so that it can reduce the tardy rate caused by the real-time messages unsuccessfully starting transmission before deadline while decreasing the average message delay at the same time. For the non-real-time class, the messages have no time constraints and the major objective is to reduce the average message delay as much as possible. Therefore, we present SPTF to achieve this aim and introduce FIFO algorithm to avoid the starvation problem caused by SPTF. Based on the simulation results, we propose an admission control algorithm to limit the number of incoming traffic flows so that the delay requirements of the admitted real-time traffic flows can be met and the network utilization is kept as high as possible.

1.3.3 Thesis Organization

In the following, Chapter 2 is the literature review on recent studies in the MAC protocols for CDMA wireless networks, followed by a detailed description of the DQRAP/CDMA protocol. In Chapter 3, the modification on DQRAP/CDMA to support real-time traffic is presented, followed by three novel scheduling algorithms that enhance the performance of the integrated traffic that is composed of real-time and non-real-time messages with certain fairness. Experimental evaluations on the performance of the proposed protocol and scheduling algorithms are made in the end of the chapter. Chapter 4 firstly describes the PDD model that can support relative differentiated service among classes and three methods to implement the model in practice. Then it gives a specific description for inter-class and intra-class schedulers. Based on the novel scheduler, a novel admission control algorithm that provides guarantee to the QoS requirements for the differentiated service in

Chapter 1 Introduction

the network is proposed. After that a series of simulations are established to evaluate the performance of the proposed scheduling and admission control algorithms. Finally, conclusions of the research work and the recommendations for further study are presented in Chapter 6.

A brief introduction of the Simulation Subroutine Set library is attached as Appendix.

Chapter 2 Literature Review

In this Chapter, we firstly introduce some recent works on the MAC protocols in wireless CDMA networks. Then we will focus on the details of the DQRAP/CDMA protocol.

2.1 MAC Protocols for Wireless CDMA Networks

The MAC protocols can be classified according to their methods of medium sharing as described in [1]. These sharing methods include *dedicated assignment*, *random access*, and *demand-based assignment*. Dedicated assignment methods allocate each user a predetermined and fixed allocation of resource, regardless of the user's need for data transmission. These kinds of schemes perform well on stream type traffic with high channel utilization. However, by a fixed assignment method, a channel is wasted whenever the user has nothing to transmit. It is widely accepted that most services in broadband environment are variable bit rate (VBR) services (e.g., bursty traffic). Services to such traffic would waste a lot of bandwidth in a dedicated assignment scheme. On the other hand, random access scheme allows all users to contend for the channel by transmitting as soon as messages available. The typical random assignment protocols like ALOHA and carrier sense multiple access with collision detection (CSMA/CD) schemes are efficient in servicing bursty traffic. Demand-based assignment schemes assign resources according to requests, or reservations, submitted by users. Demand assignments protocols typically include two stages: a reservation stage where a user requests data transmission issue, and a transmission stage where the actual data is transmitted. A reservation channel is used solely for users to request permission for data transmission.

Short reservation messages are sent to request channel access time by using some simple multiple access schemes (either dedicated or random access schemes). Once the time slots on a channel are reserved, data can be transmitted through the transmission channel. Unlike a random access protocol where collisions occur in the data transmission channel, collisions by demand assignment protocols occur only in the reservation channel. This reservation technique allows demand-assignment protocols to avoid bandwidth waste due to collisions. They are very useful for variable rate traffic and the various conditions of multimedia traffic in the integrated wireless networks. However, the additional overhead and delay caused by the reservation process can degrade the performance. Because the fixed assignment is not suitable for bursty data traffic, at the same time, it cannot support the mobility of the integrated wireless network. We just study the MAC protocols in the following three groups: pure random access, pure demand assignment, and a hybrid between these two.

The protocol proposed in [2] can be regarded as a pure random access protocol, while the protocols presented in [3-5] are combined of random access and reservation methods. In the protocols in [6-12] the mobile terminals can only obtain the transmission resources after their requests compete successfully.

2.1.1 Random Access MAC Protocols

The MAC protocol in [2] is designed for a multi-rate CDMA network to support multiple traffic types in an integrated wireless access network. It proposes a message-oriented MAC protocol that has prioritized queuing for different types of traffic. The users are

classified into types according to their traffic rate. As a message arrives from the different sources, it is buffered in a finite-length buffer for each traffic type. The user can assume three states: idle, active and blocked, based on the state of the buffer and the success of the previous transmission. If a user in idle status generates a message, the message is queued in the transmission buffer and the user assumes the active state. An active user attempts to transmit the head-of-the-queue message with probability P , which may differ in values for different queues. A lower P corresponds to a higher priority. If the transmission succeeds, the user attempts to transmit the next message in the queue if it is not empty and the user's state remains unchanged. If the successful transmission emptied the queue and no new messages have arrived, the user assumes the idle state. If the transmission fails, the user assumes the blocked state and attempts a retransmission with probability P . If the retransmission fails, the user remains at the blocked state and reattempts the transmission. An advantage of the protocol is its simplicity of implementation. It can also ensure a considerable multiplexing gain as bandwidth increases. However, the use of information messages in contention periods not only greatly increases delays due to contention resolution, but also increases the likelihood of message loss. This can adversely affect the delivery of throughput-dependent traffic (e.g., file transfers) that is sensitive to loss of information.

2.1.2 Random Reservation Protocols

The WCDMA MAC protocol for real-time and non-real-time services [3] is a proposal to be used on top of the WCDMA physical layer adopted by the Japanese Association of Radio Industries and Business (ARIB) and ETSI for their IMT-2000 standard systems (ETSI has adopted WCDMA only in the paired band or FDD mode). The physical layer is

capable of simultaneously transferring data from different traffic streams with a single remote terminal. These services can have different QoS requirements in terms of BERs, which can be met at the physical layer by applying different types of coding techniques for different types of services. The terminal has a dedicated bi-directional channel (DCH) at its disposal, where each DCH has a dedicated code. There are two basic methods for data services. Short messages can be transmitted using the random access channel (RACH), a channel common to all terminals in the cell used for issuing transmission requests as well. This mode can be used to transmit short and infrequent messages immediately with minimum overhead. In the case of larger messages, the terminal will request DCH on the RACH, i.e., a dedicated code with fast-power-controlled transmission. After evaluation, the network answers the request indicating a set of possible transmission formats (TFs) if the required resources are available. While the terminal is transmitting, the network will decrease or increase its transmission rate depending on the network load, by modifying the spreading factor of DCH code. Once the transmission is finished, the terminal will maintain the link for a certain time. If a new message is generated during that time, the mobile can start transmission immediately with the prescribed TF. If the generated message is very large, the terminal will first issue a request on the DCH asking for permission. Then, if no more messages arrive during the holding time, the link will be lost. But the terminal will keep the TFs for a certain period in order to facilitate future transmissions. For multiplexed services or multiple services transmitted from the same terminal, if the terminal wants to transmit, the network will assign a set of TF for each service. Real-time services have an allocation procedure that is very similar to the data case. The terminal will be able to use any TF for real-time

services, while the network will assign a specific TF for data services. Thus, the network has good control over the radio spectrum and efficiently uses the resource.

Another CDMA MAC protocol for integrated voice/data network is presented in [4], called Message-Reservation (for voice traffic) and Status Sensing (for data traffic) in CDMA systems (PRS²-CDMA). Uplink channels in the proposed protocol are composed of time slots with multiple spreading codes per slot based on Wideband-TD/CDMA. In the PRS²-CDMA protocol, the uplink stream is divided into fixed-length frames. The frame duration is chosen so that it is identical to the voice message generation period in voice terminals. Voice and data terminals share a group of spreading codes to transmit messages to the base station. There are three states for voice terminals, idle, reservation-contending, and reservation. But there are only two states for data terminals, idle and reservation-contending. When a new talk spurt starts, the voice terminal enters the reservation-contending state and randomly selects an available spreading code in the next slot. It transmits a reservation-request message with the transmission permission probability β . Upon capturing the preamble field of the request message, the base station immediately broadcasts an acknowledgement containing the terminal identifier and the spreading code information of the current slot. The corresponding code is reserved to the terminal and the user enters the reservation state after receiving the acknowledgement. It begins to transmit voice messages over the reserved channel of the subsequent frames until the talk spurt completion. When more than one reservation request messages collide with the same channel, the base station cannot capture the preamble field and thus does not send an acknowledgement. The collided terminals try to reserve a spreading code in the next slot. Once the base station does not detect a message from a reserved code, it regards the code

as an available code for use by other terminals in the subsequent frames. In this protocol, the data terminals do not use the reservation scheme but the spreading code status-sensing scheme. When a terminal generates a data message, it enters the contending state and randomly selects a spreading code in the current slot, which has not been already used by a voice terminal in the reservation state and not currently reserved by a voice terminal in the reservation-contending state. Upon receiving the available code status information of the current slot from the base station, a data terminal transmits a message with the permission probability β . If the base station successfully captures a data message, it also broadcasts a successful acknowledgement for the message. If the terminal receives the acknowledgement and still has messages in its buffer, it remains in the contending state and tries to transmit another message in the next slot. Otherwise, it will go to the idle state when there is no message in the buffer to transmit. If the terminal fails to transmit a message, it remains in the reservation-contending state and reattempts to send the same message in the next slot. This protocol accommodates the advantages of reservation scheme for the voice traffic, the simplicity of random access scheme for the data traffic, and the multiple access capacity of CDMA. This protocol gives a high access priority to the voice terminal. The data terminals can only access available spreading codes in the current slot that are not reserved or used by the voice terminals. Therefore, it may cause a starvation problem for the data subsystem in heavy traffic.

The MAC protocol in [5] is also a TDMA/CDMA MAC protocol. The downlink and uplink channels are separated into different frequency bands (FDD mode). The unlink channel consists of sequential frames. Each 20-ms long uplink frame is divided into multiple slots. Mobile terminals can transmit messages to the access point (AP) using

Chapter 2 Literature Review

multiple slots in each non-sequential frame. The AP can provide the various bit rates required by mobile terminals via the various combinations of the number of slots in frames and spreading codes in each frame. In each downlink time slot, the AP broadcasts uplink code status information for the next uplink time slot via an extraordinary feedback channel. The traffic service is classified into two types: connection-oriented service and connectionless message transmission service. When a mobile terminal uses connection-oriented service, it first transmits a contention message to the AP using an available spreading code that is randomly selected based on feedback information. When the AP receives a contention message from the mobile terminal, it decides whether the required channel resources can be allocated to the terminal. If there are sufficient channel resources, the AP sends an acknowledgement message, including assigned slot number and spreading codes to the mobile terminal via the feedback channel. If contention is successful and the required resources are allocated, the mobile terminal starts transmission using spreading codes in each slot and reserved slots in a frame. However, the terminal retries contention in the next slot if it does not receive a successful feedback message because of collision or insufficient channel resources. When the contention is released and the mobile terminal does not transmit any message in the reserved slot, the AP deallocates reserved slots and spreading codes. The transmission method for connectionless traffic is different from connection-oriented traffic. Upon receiving a connectionless message from an upper layer, the mobile terminal attempts to transmit messages using available spreading codes in the nearest uplink time slot. The mobile terminal can transmit more than one message using multiple available spreading codes, in contrast with conventional CDMA systems. One or more messages can be thus lost in the

case of collision, but they are retransmitted in the next slot using available spreading codes.

2.1.3 Demand Assignment MAC Protocols

The authors present a new protocol, Multidimensional Message Reservation Multiple Access (PRMA) with Prioritized Bayesian broadcast (MDPRMA BB), in [6]. This protocol is suitable for any FDMA/TDMA or TDMA/CDMA scheme. It is a derivation of the classic PRMA protocol [13] adapted to hybrid access schemes and multimedia traffic. As in PRMA, time slots of fixed length are grouped into frames, but MDPRMA BB further divides each time slot into subslots using up to eight spreading codes. These subslots are either available for contention (C subslot) or reserved for the information transfer of a particular terminal (I subslot). A mobile already registered to the network but without holding reservation will contend for a reservation using C subslots. The mobile contends in the certain C subslot with a specified probability which is service-class- and time-slot-dependent. The base station broadcasts probabilities for each type of service and each time slot of the next uplink frame in the downlink frame. The probabilities depend on the estimated number of backlogged terminals with messages to transmit in the system. They are also calculated according to protocol stability and best delay-throughput performance. Moreover, these probabilities are connected to a load-based access control to ensure control of the interference level of the CDMA component and the stability of the system. The downlink broadcast also includes the acknowledgement for successful requests in the previous frame. Usually, implicit reservation is used, that is, once a user receives the acknowledgement, it will start transmission in the C subslot that has just been used to send the request. From that moment, this C subslot becomes an I subslot and will

be reserved for that user. The reservation will last until the end of the spurt in case of voice users, or for a certain number of frames in case of data messages transmission. By controlling the contention access (transmission probabilities) and allocation (number of frames) for data services, the protocol is able to control delay requirements and dropping probabilities for different services. One of the main drawbacks of this protocol is that it is not able to support high-bit-rate data traffic or real-time traffic such as video that need multi-subslot allocations in the same frame. Additionally, it allows different services (with different BER requirements) to share the same time slot. Thus, the capacity of the slots will be variable and limited by the most demanding service (with respect to BER).

In [7], an inhibit sense multiple access-DS/CDMA (ISMA-DS/CDMA) MAC protocol is presented. When the terminal has messages to transmit, it selects one of the available codes to send a preamble to the base station. Upon receiving all the preambles, the base station assigns the spreading codes and sends the random numbers in the corresponding downlink of the selected codes. The assigned codes will be in busy state during the transmission. The mobile terminal will begin the transmission if it receives the acknowledgement correctly and the random number fits the one sent in the preamble. Otherwise, it will again attempt to access with probability p_r . If the message is not correctly sent, it will be resent in the next frame with probability p_{ret} that is inversely related to the number of retransmission. Upon completing transmission, this code will be released to a free state for other users. In DS/CDMA, users can make use of decreasing spreading factor to obtain high bit rate under low traffic load, while reducing bit rate to obtain higher spreading factor under heavy traffic load at the expense of bearing a longer delay. The paper compares the two algorithms: mobile station control (MS) Algorithm and

thresholds mobile station (ThMS) Algorithm. MS algorithm is to vary the bit rate according to the number of successful or erroneous transmissions for a given message. It simply multiplies the bit rate by two after $\min_{\text{suc}}$ consecutive successful transmissions and divides the rate by two after $\max_{\text{er}}$ consecutive erroneous transmissions. Moreover, ThMS algorithm is to vary the bit rate R_b according to the broadcast information from the base station. This decision is taken depending on the number of simultaneous codes. It defines a set of monotonically decreasing thresholds γ_i , $i=0,1,2,3,4,5$ with $\gamma_5 = 0$. Then R_b is a function of x , the number of busy codes, and the slowest bit rate v as

$$R_b = 2^i v, \quad \text{if } \gamma_{i+1} \leq x < \gamma_i \text{ for } i = 0,1,2,3,4,5$$

ThMS outperforms MS in terms of delay and throughput for both high and low traffic loads. The poor behavior of MS is mainly due to the rough estimation of the initial bit rate, which causes the algorithm to evolve until it reaches the appropriate bit rate thus leading to a decreased mean bit rate and an increased mean delay. Another important feature of ThMS is that it relays on its faster reaction capability to traffic load variations in the overall system load when compared to the MS algorithm.

Collision Resolution and Dynamic Allocation (CRDA) [8] is a slotted message-reservation multiple access technique with dedicated reservation slots. This protocol overcomes the major shortcoming of previous Message Reservation Multiple Access (PRMA) [13] schemes, i.e., contention, through the integration in the MAC protocol of a CDMA transmission mode to access the reservation slots. CRDA eliminates collisions due to possible concurrent bandwidth reservation requests made by two or more mobile stations within the same reservation slot. Uplink and downlink times are organized into

slots that are in turn grouped into TDMA frames. A message generated by voice or data terminal is partitioned to cells and one cell can be transmitted during one slot with one spreading code. Slots belonging to each uplink frame are partitioned into one request slot (R-slot) and several user information slots (I-slot), while those belonging to downlink frames are partitioned into one acknowledgement slot (A-slot) with the same number of I-slots as uplink. If a voice terminal attempts to gain an I-slot reservation, it must transmit a signature (SIG) message in a R-slot. The base station then informs the successful terminal in the A-slot paired to the R-slot in which the SIG message was transmitted. At the same time, the base station either immediately allocates an I-slot to the voice terminal if available, or queues the SIG in a first-in-first-come (FIFO) queue, delaying the allocation until I-slots available. If the base station does not receive the SIG message correctly within a predefined period, the terminal attempts to transmit the SIG in the next R-slot. After all the messages of the talk spurt are transmitted, the reserved I-slot will be released available for other talk spurt. On the other hand, the data messages are transmitted in a slightly different way from voice messages. When an incoming data message finds the local data queue not empty, the terminal sets the piggyback bit and the number of the I-slots that are needed to transmit all the messages already in the queue. If the local queue is empty, the data terminal waits for the next R-slot to send an SIG message just as for voice message. Once one I-slot has been allocated to the data message, following cells of the message will turn back to the piggybacking strategy as above. If there are still free I-slots after the requested I-slots for voice message are all assigned, the base station searches the non-empty data queues in a round robin fashion and allocates one I-slot to one non-empty queue at one time. If needed, more than one slot can also be allocated at one time to meet the bandwidth requirement. In addition, this protocol could support

various data traffic with different QoS requirements by introducing several priority levels. During the reservation phase, the queues are prioritized and the round-robin data queues are visited according to their priority level.

Wireless Multimedia Access Control Protocol with Bit Error Rate Scheduling (WISPER) protocol has been developed in [9] to take advantage of the power control characteristics of the IS-95 standard. This protocol is designed to maintain the BER of the transmission channel below a given specification. It schedules message transmission according to the BER requirements of several different traffic classes. In this protocol, the total available bandwidth is divided into two bands, one for the uplink and the other for downlink. For both bands, time is divided into frames and each frame is further divided into slots. The length of a frame is chosen to coincide with the message arrival rate of the most abundant traffic class (usually voice). For the uplink, each frame is divided into one request slot and a number of message slots. The request slot can be used for two purposes. One is to place admission requests by new remote terminals that want to be admitted to the wireless network. The other is to place transmission requests by remote terminals registered in the wireless network. Whenever a remote terminal has new messages ready for transmission, it must first send a transmission request to the base station indicating the number of messages in the new batch as well as the deadline of the messages. A remote terminal sends the transmission requests by either using the request slot or piggybacking the request in a previously transmitted data message. The latter method is used whenever possible in order to reduce contention in the request slot. In either case, requests are transmitted via an assigned primary pseudo-noise code. When a base station responds to a request for message transmission, it specifies the slot(s) and the corresponding number of

messages that can be transmitted in the next frame. To use the available bandwidth in an efficient manner, messages that have either equal or almost the same maximum BER specifications are transmitted in the same slot. In other words, the message with the most stringent BER specification determines the maximum number of messages that can be transmitted simultaneously in one slot. A novel scheduler determines the transmission order of the messages. For each new frame, the message scheduler prioritizes messages transmissions and accommodates the high-priority messages in the frame so that throughput is maximized. The transmission order is determined according to message's timeout value and the number of messages ready for transmission at each remote terminal. WISPER maintains different BERs according to the specifications of the messages in transmission. This results in an important increase in total system efficiency over those schemes that use a single BER threshold. The throughput is maximized by scheduling message transmissions according to traffic classes and the remote terminal's maximum possible transmission rate. In addition, message losses are minimized by a message prioritization scheme that determines message transmission order referring to remaining times before message timeout occurs.

The following protocols [10-12] have a close relationship with WISPER protocol. The authors present a QoS-oriented MAC protocol with fair message loss sharing (FPLS) for a hybrid TD/CDMA wireless system with messageized transmission operating in the TDD mode to accommodate asymmetric traffic in [10]. Time is also partitioned into frames of a constant duration, and each frame is further divided into time slots. Multiple accesses within each time slot are accomplished by assigning unique pseudorandom noise (PN) code sequences to each user. In the uplink, there are several access request minislots at

the beginning of the frame. The request slots are followed by a number of message transmission slots. The downlink transmission in each frame starts with a control slot that consists of a request acknowledgement subslot and a transmission permission slot. To provide each user with enough resources to satisfy its QoS requirements including the transmission BER, message loss, and delay requirements without overallocating resources, the scheduler firstly determines the priorities for the users to transmit their messages. Under high traffic, when the number of messages waiting for transmission is up to the system capacity, the priority of transmission is a function of the time out value, message loss requirements, and the traffic load characteristics. Once the priorities are determined, a message from the user with the highest priority will be first scheduled for service. Based on the transmission rate statistics and real-time traffic load information, the FPLS scheduler allocates a minimum amount of necessary resources to each user for QoS provisioning in order to acquire high resource utilization. The BER requirements are to be guaranteed by properly arranging simultaneous message transmissions and controlling their transmission power levels. The propose protocol outperforms both WISPER and discrete generalized processor sharing (GPS) scheduling schemes in terms of resource utilization. Furthermore, it reduces the computational complexity greatly by assigning the priority to each user instead of each individual message in the scheduling.

Another derived study from WISPER is presented in [11]. In WISPER, the authors suggest avoiding scheduling simultaneous transmissions of messages with different BER specifications to increase the efficiency of MAC protocols for multi-code CDMA systems. This observation is the primary motivation in this paper. In a CDMA system, there are multiple simultaneous spreading code transmissions over the same time interval.

Then the wireless channel can be regarded as multiple servers in which the number of servers varies with time. Each server provides an identical service rate (i.e., a fixed transmission bandwidth), and the service quality (i.e., transmission BER) deteriorates uniformly as the number of servers increases. There is a single scheduler running in this system to achieve QoS requirements, and all the servers are governed by this scheduler. The authors present that among all multi-server scheduling policies with “weak load balance” property, the largest delay schedulable region is achieved when we use non-preemptive earliest-deadline-first (EDF) scheduling policy and normalize all message sizes with respect to the worst transmission BER. If a message is chosen to serve, this message is divided into a specified number of equal-size parts and transmitted with the corresponding number of spreading codes in parallel. This can ensure that real-time data simultaneously transmitted always have equal BER requirements since they actually come from the same message while WISPER transmits messages with equal or similar BER requirements in the same time slot. Nevertheless, this protocol assumes that the wireless channel has stable transmission bandwidth. It needs further work to become robust enough to combat highly dynamic bandwidth variation and time-varying BER characteristic due to severe multipath, spatially and temporally varying fading effects of wireless channels.

Power levels of all code channels are assumed equal in WISPER, i.e., the power levels are not minimized. In addition, there is no solution in WISPER to address the call admission control (CAC) issue. Thus, the author in [12] proposes a new combined message scheduler consists of the message prioritizer, the joint rate- BER-scheduling scheme, and the resource controller based on minimum-power allocation. It is proved that the

minimum level of a code channel can be achieved when the signal-noise-ratio (SNR) of one mobile terminal of a specific service equals to the target SNR value. The prioritizer determines which batch of messages is selected according to QoS requirements. In order to reduce message delay and message loss ratio, the priority of a batch needs to be proportional to the number of remaining messages in the batch and inversely proportional to its remaining time before timeout. After the prioritizer selects a batch, the joint rate-BER-scheduling scheme allocates the appropriate time slots and code channels to accommodate messages in this batch. The scheme works iteratively until all messages in the batch have been accommodated or no code channel is available for the batch in the current frame. If no code channel is available and there are still messages remaining in the batch, these messages are served in the next frame when they are not out of time. The resource controller controls the capacity of each time slot. After scheduling is completed for the entire frame, the minimum-power allocation algorithm is used to allocate the received power level to each code channel in each time slot. The proposed MAC protocol outperforms WISPER because the minimum power allocation and joint rate-BER-scheduling can reduce capacity consumption in each time slot. The CAC algorithm based on minimum power allocation enhances the performance of the MAC protocol in terms of message loss ratio and message delay although the system throughput is reduced, which is rather small compared to the significant improvement in performance.

The random access method used in [2] allows the users to contend the resource as long as there are free channels in the system. The transmission begins when competition successful, which may result in zero waiting time in the network. When the traffic load is heavy, however, the competition becomes drastic and more collisions may happen that

leads the users to attempt iteratively for resource access. Thus, it incurs uncontrollable delay in the network before successful transmission. In this way, the base station operates no better than a receiver does with multiple inputs from the terminals. It cannot control the collision resolution and resource allocation among the users. The hybrids in [3-5] combine the random access and demand-based assignment schemes to support the integrated traffic in the system. The demand assignment methods allocate resource to the connection-oriented traffic, i.e., long voice messages, according to transmission request. The random access schemes provide service to connectionless traffic, i.e., short bursty data messages. There is a main shortcoming for these hybrid random access/demand-based assignment protocols. The connection-oriented traffic is generally allocated higher priority to service than connectionless traffic. This unfairness between them may lead to starvation problem to the connectionless traffic. Thus, we can only provide best-effort service to them without any QoS guarantee.

In the demand-based assignment protocols in [6-12], each link frame structures are divided into two parts, reservation phase and transmission phase. The proportion of reservation phase in the entire frame period is generally much smaller than transmission phase. The request transmission and the data transmission work in parallel and have no effect on each other. In other words, one user acquires the service at least after a delay that is the sum of the request transmission time in the uplink channel and the acknowledgement transmission time in the downlink channel, where we assume that the base station can deal with the request in such a short time that can be ignored and at that time there is still free resource in the system to be allocated to the user.

Demand-based channels are useful for multimedia traffic, i.e., variable-bit-rate traffic. The controllability on the QoS provisioning is so attractive for the MAC protocol design that the performance depravation, addititheonal overhead and delay, caused by the reservation process are tolerable. Therefore, our research work concentrates on demand-based assignment protocols to provide service to various traffic types in CDMA wireless network.

2.2 A Near-Optimum MAC Protocol Based on DQRAP for a CDMA Mobile Communication System

2.2.1 System Background

One widely studied reservation-based MAC protocol, the distributed queue request update multiple access (DQRUAM), has been reported to effectively support data traffic in [14]. This protocol uses a certain number of access minislots for reservation purposes. A mobile node with data to transmit will send an access request in one of these minislots by a slotted-ALOHA strategy. The request contains the identification number of the user and the type and quality of the demanded service. In general, merely using control minislots makes the system more complex as it is necessary to have time slots with different sizes. It is observed that all existing tree protocols without minislots use data slots to resolve collisions and thus lose the channel capacity of all the empty slots or collided messages. The suggested improvements to tree protocol seek to reduce the number of collisions and empty slots. But they cannot eliminate this type of efficiency loss. In the DQRUAM protocol, the time axis is divided into slots, each of which has two fields. The first field is the access field, which is further divided into three control minislots. The second field is the data field, where users will transmit their messages. It is assumed that every user has

perfect slot and minislot synchronization. Then the central controller could schedule and control data transmission according to the network condition and quality of service of the traffic at different users. Another example is the distributed queuing random access protocol (DQRAP) [15-16]. This protocol also uses three control minislots based on a tree-type collision resolution algorithm. It was initially designed for a TDMA system. Its performance can approach that of an ideal M/D/1 queue, reaching maximum stable throughputs close to one, and maintaining its stability for traffic loads up to channel capacity. An improvement can be found in [17], where a near-optimum MAC protocol based on DQRAP for CDMA wireless network has been proposed, named as DQRAP/CDMA. The terminals are able to change the spreading code for data and request transmission on a slot-by-slot basis. The messages generated by one terminal are split into slot-duration packets and put into a buffer. Each message will be sent with the same spreading code. But not all the packets pertaining to one message will necessarily be sent with the same spreading code. This protocol can behave like a random access protocol when the traffic load is light and switch automatically to a reservation protocol when the traffic load increases. DQRAP/CDMA can achieve a very good performance in supporting data traffic. In addition, its implementation is much easier than that of the DQRUAM algorithm.

The DQRAP/CDMA protocol uses two concatenated distributed queues: the collision resolution queue and the data transmission queue. Once a terminal generates a message, it selects a spreading code and sends a request in one of the control minislots pertaining to this code following a certain set of rules. If the terminal fails to acquire resource (i.e., the request collides with one or more requests from other messages), it enters the collision

resolution queue. Collisions are then resolved in the order fixed by the queue discipline. In addition, the data transmission queue contains the terminals that have succeeded in their request and are waiting to send messages to the base station also following the order fixed by the corresponding queue discipline. Collision resolution and data transmission processes work in parallel.

All the terminals must have four integer counters, which represent the two logical distributed queues. They are denoted as TQ , RQ , pTQ , and pRQ . TQ is the number of terminals waiting for service in the distributed transmission queue. RQ is the number of collisions waiting for resolution in the distributed collision resolution queue. pTQ is the position of a given terminal in the data transmission queue, and pRQ is the position of that terminal in the collision resolution queue. The value of pTQ (or pRQ) from 0, meaning that the terminal does not have any position in the corresponding queue, to TQ (or RQ), 1 being the first position of the queue. TQ and RQ have the same value for all the terminals in the system (i.e., they represent distributed queues), while pTQ and pRQ have a specific value for each terminal. Both queues are assumed to be in First-In-First-Out (FIFO) order. All four values are initially set to zero and must be kept updated with the feedback information sent by the base station, each slot, via a broadcast channel, and following a set of rules. It consists of ternary state data for each control minislot of every spreading code, and has to include a final-packet-bit for each code.

The three different states that the base station must be able to distinguish are: empty, success, and collision. A collision will occur when more than one terminals transmit requests in the same minislot of the same spreading code. The final-packet-bit is the mark

that all the terminals must send when they are transmitting the last packet from one message. This flag bit must be ON in the last packet of each message, and must be OFF in all the other packets. This mechanism allows all packets from a message to be transmitted with a single request and minimizes the delay jitter between these packets. Nevertheless, if propagation delay in the system prevents the terminals from receiving the feedback information about this final-packet-bit before the next data slot begins, an empty slot loss is produced at the end of each message. If messages are known to be short (for example, ATM cells), it should be possible and convenient to switch off this mechanism and consider all packets formed by a single message.

2.2.2 Protocol Description

As the original DQRAP, the new protocol algorithm consists of three sets of rules, DTR (Data Transmission Rules), RTR (Request Transmission Rules), and QDR (Queuing Discipline Rules) that each terminal has to follow at the end of each slot. However, the order in which these rules of DQRAP/CDMA are executed is QDR, DTR, and RTR, different from that of the original DQRAP protocol. Let us consider N data terminals that share a CDMA channel with K available spreading codes to communicate with a base station. Here we assume that before the rules are executed at the end of each slot, the terminal has already received the feedback information correctly from the base station about the state of the control minislots of the previous slot. Some rules have initial conditions that must be true to execute the corresponding operations. If the assertion is not verified, the algorithm simply jumps to the next rule. When all the rules have been checked, the current slot finishes and a new one starts.

1) QDR

- a) Each terminal increments the value of TQ by one unit for each control minislot in the success state, taking into account the feedback information for all control minislots from any of the K spreading codes.
- b) Each terminal reduces the value of TQ by one unit for each message correctly received by the base station with the final-packet-bit set to ON (the final-packet-bit).
- c) If $RQ > 0$, each terminal reduces the value of RQ by $\min(RQ, K)$ units.
- d) Each terminal increments the value of RQ by one unit for each control minislot in the collision state, taking into account all the minislots from the K spreading codes.
- e) Depending on its state, and the results of the control minislots, each station calculates the value for its own pTQ and pRQ . That is, if it has sent a request and this request has succeeded, it calculates its position among all the succeeding minislots and sets pTQ to the corresponding value at the end of TQ . For this purpose, all the successes are sorted using the order of spreading code to which they belong, and within the same spreading code, using a time arrival criterion. On the other hand, if the request has collided with others, the terminal calculates its position among all the present collisions and sets pRQ to the corresponding value at the end of RQ . If it has not sent any request, pTQ and pRQ follow the same update rules as TQ and RQ , respectively, but only if the initial values are other than zero.

2) DTR

- a) Each terminal increments the value of TQ by one unit for each control minislot in the success state, taking into account the feedback information for all control minislots from any of the K spreading codes.
- b) If a terminal has $pTQ > 0$ and $pRQ \leq K$, it transmits the first packet of its buffer using the spreading code K_{pTQ} . If this packet is the last one of the current message, the station sets the final-packet-bit to ON.

3) RTR

- a) If $RQ < K$, each terminal that has $pRQ = 0$ and $pTQ = 0$ and data ready for transmission randomly selects one of the control minislots of the spreading code K_{RQ+1} and transmits a request in it.
- b) If a terminal has $pRQ < 0$ and $pTQ < 0$, it randomly selects one of the control minislots of the spreading code K_{pRQ} and transmits a request in it.

According to DTR-1, it will allow the newly arrival messages to be transmitted without sending requests when the channels are not all busy, thus the protocol is a free random access protocol when the traffic load is light. However, when the newly arrival messages are more than one, they will collide when occupying the same code channel. Then the corresponding terminals will send request to contend in minislots, and the protocol switches smoothly and automatically to a reservation protocol when traffic load becomes heavy.

Shown in Figure 2.1, the DQRAP/CDMA protocol can be modeled as a queuing system that consists of two subsystems: the collision resolution subsystem and the transmission

subsystem, very similarly as the original DQRAP protocol. Both subsystems have as many servers as available spreading codes (i.e., K). The elements in the system are the messages generated by the users, although they only use the control minislots for accessing purpose in the collision resolution subsystem. The feedback line in the subsystem represents that the messages that collide in their requests must enter the queue again until they succeed.

Figure 2.1 Model of DQRAP/CDMA.

Chapter 3 Providing Integrated Service in CDMA Wireless Networks

3.1 Motivation

In current wireless communication networks, different types of traffic could exist in one system and share the limited bandwidth of a communication channel and other resources. The wireless network should have the capacity to provide flexibility and efficiency to allow the existence of diversities of traffic, such as integrated traffic and multimedia traffic.

Integrated traffic is composed of real-time streams, in which messages have time constraints, and non-real-time streams without time constraints. The most important aspect of the time-constrained applications is that a message generated at a source station must be received at its destination station within a given amount of time. This time period is referred as a message laxity. If the delay to a message in the network exceeds its time constraint, the message is considered as late. There are two types of real-time messages. The hard real-time message is one kind of message that its time to be delayed in the network is not allowed to be longer than the message laxity. The potential late message must be dropped rather than transmitted. The soft real-time message is another kind of message that it is tolerable to delay later than the message deadline. A scheduling algorithm for a MAC protocol that serves real-time messages is to transmit the messages to meet their time constraints as much as possible. However, the service that benefits the

Chapter 3 Providing Integrated Service in CDMA Wireless Networks

real-time traffic may cause unfairness to the non-real-time streams of the integrated traffic. The challenge to the design of scheduling algorithms is how to serve the real-time messages without violating their deadlines while providing transmission service to the non-real-time messages with reasonable performance in order to reach some extends of fairness at the same time.

The DQRAP/CDMA protocol can achieve a very good performance in supporting best-effort data traffic. However, it only considers the scheduling of unique type of messages in the network. Although it has the potential to support real-time traffic in a CDMA-based wireless network, it needs modifications in the scheduling algorithms to efficiently reduce the waiting time of the real-time messages in the queue. Furthermore, the protocol should provide service to integrated traffic that needs some fairness between real-time and non-real-time traffic, which reflects the trends that future wireless networks should simultaneously accommodate various kinds of data traffic in one network plane. A novel MAC protocol has been proposed in this project, which is named as distributed queuing random access with assignment protocol (DQRA+AP/CDMA). It also presents the corresponding scheduling algorithms to provide integrated service to the mixed applications, which is composed of real-time and non-real-time streams in a CDMA-based wireless network. DQRA+AP/CDMA is designed based on the DQRAP/CDMA protocol with consideration of real-time service. The major contributions in this study are to 1) combine the random access scheme with the fixed assignment method to solve the collisions among the transmission requests when they access the same control subslots; 2) introduce a time-based priority scheme to enhance real-time service to the time-constrained messages; 3) propose a simple scheduling scheme to support real-time and

non-real-time traffic at same time; 4) introduce a cost-based scheduling algorithm to enhance the performance of the integrated service with certain fairness.

3.2 System Model and Service

It is assumed that there are two kinds of traffic, real-time and non-real-time messages, integrated in the network. Each real-time message has its time constraint represented by laxity, which is time duration from the message enters the network to the time this message should begin transmission, while the non-real-time messages have no such delay requirements.

As the original DQRAP/CDMA protocol, there are two distributed queues residing at each mobile node. One is called as the collision resolution queue. The other is the data transmission queue. Each node has four variables to indicate the network situation regarding the two distributed queues. One of them is TQ , which is the number of messages waiting for transmission in the distributed data transmission queue. Another is RQ , which is the number of collisions waiting for resolution in the distributed collision resolution queue. The third is pTQ , which is the position of the corresponding mobile node in the data transmission queue. At last, pRQ is the position of that mobile node in the collision resolution queue. The values of the last two variables can be 0, meaning that a mobile node does not have any position in the corresponding queues, which further means that the mobile node has not any transmission request currently. The value of pTQ (pRQ) can range from 0 to TQ (or RQ) respectively. TQ has the same value for all the mobile nodes as similar as RQ , while pTQ and pRQ have specific values for each mobile node. All four variables are initially set to zero and must be kept updated with the

Chapter 3 Providing Integrated Service in CDMA Wireless Networks

feedback information from the base station through the downlink channel. The network architecture can also be modeled as a queuing system that consists of two subsystems: the collision resolution subsystem and the transmission subsystem. The system model of the RQ and TQ can be expressed in Figure 3.1.

Figure 3.1 System Model of Distributed Queues

The network service can be described as a MAC protocol, in general. The proposed DQRA+AP/CDMA protocol, which evolves from the DQRAP/CDMA protocol, can be presented as follows. When a message is generated at a mobile node, the node will first randomly select a spreading code as a transmission channel and send a request to indicate its data transmission in one of the control minislots pertaining to this channel. The transmission of the requests in the control minislots will be detected by the base station. Then the base station will broadcast the feedback information on the states of the control minislots to mobile nodes in the downlink channel, before next time slot begins. If it is indicated that a request collides with one or more requests from other mobile nodes, all the collided nodes will enter the collision queue in a certain order waiting for collision resolution. On the contrary, if a request has been successfully transmitted, the

Chapter 3 Providing Integrated Service in CDMA Wireless Networks

corresponding node will enter the data transmission queue and wait to send data to the base station.

Once a collided node has entered the collision resolution queue, it will not experience a competition for a control minislot any more. Rather it will be directly assigned into a control minislot in the next or following time slots. Thus, the collision will be solved quickly. Collision resolution and data transmission processes can work in parallel. Each message in the data transmission queue will be transmitted through any logical channel once enough data transmission slots for the message of a certain channel are available. One message will be broken into packets for transmission as the data transmission slot duration is defined as long as the transmission time of one packet. However, before the transmission, a message with time constraint will be checked to tell whether it has been delayed in the queues longer than its time constraint. If yes, the message will be dropped. Otherwise, it can be successfully transmitted to the base station. During the collision resolution and transmission period, only the mobile nodes, which are not in either the collision resolution queue or the data transmission queue, can send requests to the base station for data transmission so that the total amount of traffic could be controlled to make the network stable.

3.3 Scheduling Issue and Algorithms

From the description on the proposed DQRA+AP/CDMA protocol in the previous section, it is obvious that there are scheduling issues involved in both the collision resolution and the data transmission procedures. In order to efficiently solve the request

Chapter 3 Providing Integrated Service in CDMA Wireless Networks

collisions and provide integrated service to different types of traffic with good performance, careful scheduling is required. The function of scheduling algorithms can affect the performance of the network services. In this section, the proposed collision resolution scheme will be discussed in detail and several scheduling algorithms have been proposed for the DQRA+AP/CDMA protocol to support integrated service in the specified CDMA wireless networks.

3.3.1 Collision Resolution Scheme

The collision resolution scheme is developed based on the operation of the DQRAP/CDMA protocol. It is noticed that by the DQRAP/CDMA protocol, once there are more than one transmission requests from multiple mobile nodes simultaneously competing the same control minislot, the collision will happen. Those collided requests will enter the collision resolution queue in FIFO order waiting for the following time slots to compete for other minislots. It is possible for some transmission requests in the collision resolution queue to experience more than one round of collisions that results in the messages being delayed quite a long time at the mobile nodes. It is not tolerable for the real-time traffic since each message in the real-time traffic has a time constraint that cannot be expired. In order to make the protocol to serve the urgent messages quickly, we have an idea to combine the random access method with the fixed assignment approach to arrange the collided requests out of the collision resolution queue as fast as possible. Three approaches to handle the collision resolution queue have been proposed. First, each collided transmission request in the collision resolution queue will be assigned a dedicated control minislot in the following time slots. Each logical channel will reserve a control minislot for the fixed-assignment of one collided request while the remaining $m-1$ control minislots are still available for the newly incoming transmission requests to

Chapter 3 Providing Integrated Service in CDMA Wireless Networks

compete. In this way, each collided request has no chance to experience more than one round of collisions so that it should stay in the collision resolution queue for a short time period. The second step is to assign the control minislots to the collided requests in an order referring to their time constraints. A parameter is defined to present the message time constraint, which is the message laxity. The messages with shorter laxities will get their corresponding transmission requests to be assigned to control minislots earlier so that the transmission requests of the urgent messages will not be delayed in the collision resolution queue longer. In order to prevent scheduling starvation, we have the third policy that the transmission requests of the messages with earlier arrival time will be considered for the assignment sooner. By this method, some collided transmission requests that cannot have a control minislot successfully in the next time slot will definitely get their control minislots in the exact following time slots. The entire scheme on the collision resolution queue could be expressed as follows. The scheduler will first examine the arrival time of each collided transmission request in the queue at the beginning of every time slot. The queue will then be kept in a FIFO order. However, among the transmission requests that have been collided in same time slot, those with shorter laxities will be assigned control minislots earlier.

3.3.2 Data Transmission Queue Scheduling

How to handle the data transmission queue is another scheduling issue in the DQRA+AP/CDMA protocol. In order to meet the time constraints of real-time messages as much as possible, we hold the principle to schedule the transmission of the more urgent messages earlier. At the same time, we have to prevent the messages with loose time constraints from being delayed too long in the network. We propose some scheduling algorithms to control the data transmission queue.

3.3.2.1 Most Urgent First (MUF) Scheduling for Real-time Traffic

In order to benefit the real-time traffic, a time-based priority scheduling scheme is designed. The idea to schedule the transmission of the real-time messages in the queue is that the one with the smallest difference between its laxity and experienced waiting time will be transmitted earliest. The waiting time of a message is defined as the period since the message is generated at the mobile terminal until the transmission begins. By this scheme, the messages with tighter time-constraints and longer delays will be considered transmitting ahead of others. It evaluates the priority of message i in the data transmission queue by the following formula: $P^i = d^i - w^i$, where P^i is its priority, d^i is the time constraint or laxity of message i , and w^i is the delay that message i has experienced. By this priority scheme, the messages with the less value of P^i will be assigned higher priorities for transmission. In other words, more urgent the message, earlier its transmission begins. This scheduling algorithm is named as the most urgency first scheduling (MUF).

The time-based scheduling scheme MUF will benefit the real-time traffic in terms of low message loss rate due to deadline expiry. However, it can also cause longer delays to the messages with larger laxities. Thus, we modify the MUF scheduling algorithm with the goal to lower the average message delay with still low message loss rate. The modification on the scheduling is that in a relatively long term, the transmissions of the messages are scheduled in FIFO way, while in a short term, the real-time message transmissions are arranged in MUF order. The new algorithm is named as the most urgency first with first-in-first-out scheduling (MUF-FIFO). It is expected that with

Chapter 3 Providing Integrated Service in CDMA Wireless Networks

MUF-FIFO scheduling, a relatively low message loss rate as well as a low average message delay could be achieved when the network only serves the real-time traffic.

3.3.2.2 MUF-FIFO Scheduling for Integrated Traffic

However, it may cause starvation of the non-real-time traffic transmission in a network with an integrated traffic, which will result in a worse performance of the network in terms of higher message delay. MUF-FIFO is extended to introduce fairness for non-real-time traffic so that the starvation problem for them can be solved. To implement the scheduling scheme, we further split the data transmission queue into two subqueues: one for real-time and the other for non-real-time traffic. The messages in the data transmission queue enter the two subqueues according to their types respectively and wait for service. Let $D_i(t)$ be the sequence of messages arriving at subqueue i during every t length period, thus the incoming messages are grouped in batches. As described in MUF-FIFO, the real-time messages in an earlier batch should be delivered before those in a later batch. So do non-real-time messages. This ensures the message to be transmitted in FIFO order in a long term. For the real-time and non-real-time messages arriving in the same batch, the scheduler provides a higher priority to real-time traffic. So all the real-time messages should be delivered before those non-real-time messages in the same batch. Therefore, all the real-time messages in a batch $D_{RT}(t)$ are first transmitted in MUF order followed by the transmission of the non-real-time messages in the same batch $D_{NRT}(t)$ in FIFO order. We expect this scheduling algorithm to support a more reasonable balance on the performance of the network under an integrated traffic than that in MUF.

3.3.2.3 MUF with Comparison (MUFC) Scheduling for Integrated Traffic

Based on MUF-FIFO, we further propose another scheduling algorithm named as the Most Urgency First scheduling with Comparison (MUFC). This scheme is to schedule the real-time transmission as soon as possible while arranging the non-real-time transmission without violating the time constraints of real-time messages so much. Similar as in MUF-FIFO, all the messages in an earlier batch are served before those in a later batch, and real-time messages in the same batch are scheduled in MUF order. It is modified that the non-real-time messages are delivered in Smallest Processing Time First (SPTF) order but not in FIFO order as MUF-FIFO. Thus, the head message in the non-real-time subqueue is the shortest one among all non-real-time messages in the same batch. Another change is that the scheduler does not provide an absolutely higher priority for real-time traffic than non-real-time traffic in the same batch. It inserts the non-real-time messages into the process queue of real-time messages according to a particular comparison result. As described in MUF-FIFO, the priority of message i in the real-time data transmission subqueue can be determined by the following formula: $P^i = d^i - w^i$, where P^i , d^i , and w^i are respectively its priority, laxity, and the delay that message already experienced in the system. The non-real-time messages will be kept in the non-real-time data transmission subqueue, in which the messages will be in an order that the message with shorter transmission time will be sent earlier. The priority of message i in the non-real-time data transmission subqueue is only determined by its transmission time, $P^i = M^i$, where P^i and M^i are respectively the message's priority and transmission time. By the MUFC scheme, the decision on the message transmission is made by comparing the features of the head messages of the real-time and non-real-time data transmission subqueues. Let M^n be the transmission time of the message n at the head of the non-real-time data transmission

Chapter 3 Providing Integrated Service in CDMA Wireless Networks

subqueue, d^r be the laxity of the message r at the head of the real-time data transmission subqueue and w^r be its delay in the system. The decision depends on the comparison as follows: If $d^r \geq w^r + M^n$, the non-real-time message n will be transmitted first, followed by the real-time message r . Otherwise, the real-time message r will be transmitted only. However, at most one non-real-message is permitted to transmit between the deliveries of two consecutive real-time messages. Thus it reduces the degree that the real-time messages transmissions are postponed by the non-real-time messages. By the MUFC scheduling, certain degree of fairness to the transmission service to both real-time and non-real-time traffic streams could be expected.

3.3.2.4 Cost-Based Priority Assignment Scheduling for Integrated Traffic

A dynamic-priority scheduling scheme is our third proposal for the data transmission scheduling to enhance the service performance of the integrated traffic in the wireless network, which is named as Cost-Based Priority Assignment (CBPA). It determines the relative service positions of the real-time and non-real-time messages in the same batch in a more delicate way than MUFC. Notice that a message will occupy more of the network resources if it stays in the network longer. In other words, its transmission cost will increase as the message waiting time in the system. We refer the message length and its relative laxity as the factors to the transmission cost. The following formula is to assign the priority to the messages i : $P^i = \alpha \tilde{d}^i + \beta \tilde{M}^i$, where P^i is the dynamic priority assigned to message i , $\tilde{d}^i$ is its normalized relative laxity that represents the message urgency, $\tilde{M}^i$ is its normalized transmission time, α is the coefficient for the relative laxity, and β is the coefficient for the message transmission time. In the CBPA scheme, the message with the

Chapter 3 Providing Integrated Service in CDMA Wireless Networks

smallest value of P_i will be allocated the highest priority, which takes the least cost of the network resources. In order to calculate the normalized values of message transmission time and the real-time message laxities, it is necessary to work out their maximum values first. They are random variables following exponential distributions with different mean values. The 95% data of the probability density function is assumed for simplicity. Studies by Ma and Huang [20] show that for the message transmission time as a random variable, M , following an exponential distribution with mean value M_{avg} , the maximum value of that variable can be approximated as $M_{max} = 3 \times M_{avg}$. Similarly, for the laxity as a random variable, L , its maximum can be approximated as $L_{max} = 3 \times L_{avg}$. $\tilde{d}^i$ is 1.0 for non-real-time messages. Both of the coefficients α and β are selected as 1.0 in the formula. But they can be also set to be other values. If $\alpha > \beta$, the priority assignment is preferred to real-time traffic. Otherwise, it is preferred to non-real-time traffic. The setting of the coefficients makes the CBPA more flexible and scalable. Since the dynamic priority has been worked out referring to two features of the messages jointly, it is quantitatively more accurate to determine the message service order. Then a better performance of the integrated service supported by the wireless network could be expected.

3.4 Performance Evaluation

We have performed a series of simulation experiments to evaluate the performance of the proposed DQRA+AP/CDMA protocol with the proposed scheduling algorithms. With the DQRA+AP/CDMA protocol, we compare the performance of MUF, MUF-FIFO, MUFC, and CBPA. The simulation results show that the DQRAP/CDMA with MUF scheduling can provide better service in terms of lower average message loss rate to real-time traffic

Chapter 3 Providing Integrated Service in CDMA Wireless Networks

than the original DQRAP/CDMA protocol, while the former suffers a higher average message delay than the latter. The performance of the scheduling algorithms, MUF-FIFO, MUFC, and CBPA, are all further more desirable than the original DQRAP/CDMA protocol, not only in terms of lower average loss rate but also shorter average delay for real-time traffic. Moreover, the three algorithms can provide fairness to non-real-time traffic and serve them with reasonable performance.

3.4.1 Experimental Design

In order to validate the operations and evaluate the performance of the proposed scheduling algorithms, discrete event simulation experiments have been conducted. A perfect minislot state detection is assumed in the experiments to focus on the silent feature on the scheduling issues and algorithms. We compare the simulation results on the performance among three different scheduling algorithms with the DQRA+AP/CDMA protocol to demonstrate the performance improvements. The entire network is assumed to have $N=100$ mobiles and $K=16$ CDMA spreading codes as logical transmission channels. The messages are generated in a Poisson process. The message length follows exponential distribution with average message length of 6400 bits. The real-time message laxity also follows exponential distribution with average value of three times the average message length. A lossless, work-conserving, and non-preemptive link with capacity 640 of bits/slot is supposed. There are $m=3$ access control minislots for each data transmission slot on each logical channel.

The following three flow charts Figures 3.2 – 3.4 describe the procedures of the DQRA+AP/CDMA with the different scheduling algorithms in the simulation program. The whole process of data message service for a mobile from the data generation till the

Chapter 3 Providing Integrated Service in CDMA Wireless Networks

whole message is successfully received by the base station can be divided into three parts. Firstly, the mobile should send request to the base station for channel competition. After receiving the request, the base station has to determine which request is successful and how the corresponding mobile is scheduled in the queue waiting for service. When there is a channel available, the appropriate mobile is allowed to begin the data transmission. There are two major queues in the system, request queue and data transmission queue, as described above. In DQRA+AP/CDMA protocol, the successful mobile will enter the data transmission queue waiting for service while the failed mobile will enter the request queue and some head mobiles in the request queue will directly enter the data transmission queue without retransmitting the request in the next time slot. Additionally, there are two other intermediate queues for those mobiles to wait in before they enter to the data transmission queue. One is for real-time mobiles which have real-time messages to transmit, and the other is for non-real-time ones. The real-time and non-real-time mobiles are scheduled in the corresponding queues according to different scheduling algorithms. When the data transmission queue is emptied, or there are no mobiles waiting in the queue, all mobiles in the two intermediate queues will enter the data transmission queue. When there is a channel available, the head mobile in the transmission queue will be selected to service. It is clear that the discrimination among the three simulation programs is the part of the scheduling of the real-time and non-real-time messages waiting for service in the intermediate queues. Other procedures before and after this one are all the same in the three simulation programs. For convenience, it assumes that the length of the message of the head mobile in the non-real-time intermediate queue is $l^{head-nrt}$, the length of the message of the head mobile in the real-time intermediate queue

Chapter 3 Providing Integrated Service in CDMA Wireless Networks

is $l^{head-rt}$, and the deadlines of the first two real-time mobiles' messages are D^1 and D^2 , respectively.

Chapter 3 Providing Integrated Service in CDMA Wireless Networks

Figure 3.2 Flow Chart of Simulation Program for DQRA+AP/CDMA/MUF+FIFO

Chapter 3 Providing Integrated Service in CDMA Wireless Networks

Figure 3.3 Flow Chart of Simulation Program for DQRA+AP/CDMA/MUFC

Chapter 3 Providing Integrated Service in CDMA Wireless Networks

Figure 3.4 Flow Chart of Simulation Program for DQRA+AP/CDMA/CBPA

3.4.2 Experiment Results

We compare the original DQRAP/CDMA protocol with the DQRA+AP/CDMA protocol with MUF and MUF-FIFO at first. The corresponding simulation results are presented in two groups of figures. In the first group, Figure 3.5 and Figure 3.6 present the simulation results on the performance of the protocols with the corresponding scheduling algorithms under unique real-time traffic. In the second group of rest figures, Figure 3.4-3.8 present the performance of the protocols under an integrated traffic with a mixture of certain percentages of real-time traffic and non-real-time traffic. Then we compare the performance of the three algorithms, MUF, MUFC, and CBPA, supporting the integrated traffic composed of real-time and non-real-time messages. The major metric for evaluating real-time service is the message loss rate, while the major metric for evaluating service is the average message delay. Figures 3.9-3.14 show the comparison results for this part. Finally, Figure 3.15 depicts the performance of the CBPA algorithm when the amount of the real-time traffic changes.

Figure 3.5 shows the relationship between the average message loss rates and the traffic load in terms of the number of messages per time slot under only real-time traffic. It is obvious that the average message loss rate of the original DQRA+AP/CDMA protocol with MUF scheduling is much lower than that of the DQRAP/CDMA protocol, in which time constraints of messages have not been considered for the transmission scheduling. When the traffic load is heavy, the average message loss rate by the DQRA+AP/CDMA protocol with MUF scheduling is only around 60% of that by the DQRAP/CDMA protocol. The reason is evident that the DQRA+AP/CDMA protocol with MUF scheme schedules the transmission of real-time messages according to their time constraints. It

Chapter 3 Providing Integrated Service in CDMA Wireless Networks

results in that more urgent messages can be transmitted earlier and fewer messages will be lost due to the violation of their time constraints. It is also clear that the average message loss rate by the DQRA+AP/CDMA protocol with MUF-FIFO scheduling is higher than that by the DQRA+AP/CDMA protocol with MUF scheduling. However, the latter has still achieved more than 15% improvement over that by the DQRAP/CDMA protocol. Its performance in terms of message loss rate is moderate and reasonable.

Figure 3.5 Average Real-time Message Loss Rate vs. Total Traffic Load

Figure 3.6 shows the relationship between average message delay and the traffic load under unique real-time traffic. It is clear that the average message delay by the DQRA+AP/CDMA protocol with MUF scheduling is much higher than that by the DQRAP/CDMA protocol under heavy traffic load. It has performed worst in terms of average message delay although it can achieve the lowest message loss rate. Since this scheduling algorithm always arranges the most urgent message to be transmitted first, fewer messages are dropped because of violating the deadline. At same time, the messages with larger laxities will be delayed longer for transmission. However, when the traffic load is not heavy, the DQRA+AP/CDMA protocol with MUF scheduling can still

Chapter 3 Providing Integrated Service in CDMA Wireless Networks

achieve some improvement over the DQRAP/CDMA protocol because of the quick collision resolution scheme of the DQRA+AP/CDMA protocol. The DQRA+AP/CDMA protocol with MUF-FIFO scheduling performs almost same as the DQRAP/CDMA protocol as the traffic load increases. It also has some improvement over the DQRAP/CDMA protocol when the traffic load is moderate because of the quick collision resolution scheme.

Figure 3.6 Average Real-time Message Delay vs. Total Traffic Load

From the above two figures, it is obvious that the DQRA+AP/CDMA protocol with MUF-FIFO scheduling can achieve moderate improvement over the DQRAP/CDMA protocol in terms of the message loss rate while it almost keeps the advantage of the DQRAP/CDMA protocol in terms of the average message delay. The DQRA+AP/CDMA protocol with MUF scheduling can outperform much more than the DQRAP/CDMA protocol in terms of message loss rate. However, its performance in terms of average message delay is worst when the traffic load is heavy. Based on its extreme performance, it can be predicted that the DQRA+AP/CDMA protocol with MUF scheduling could not benefit the non-real-time messages under an integrated traffic environment. Rather it will

Chapter 3 Providing Integrated Service in CDMA Wireless Networks

serve the non-real-time message transmission with worse performance. On the contrary, the DQRA+AP/CDMA protocol with MUF-FIFO scheduling can be able to serve the integrated traffic with reasonable performance.

Figure 3.7 shows the relationship between the average message loss rates and the traffic load under an integrated traffic load. One half of the traffic is real-time while the other is non-real-time. The DQRA+AP/CDMA protocol with MUF scheduling has the best performance. However, it has the worst performance in terms of average message delay as predicted.

Figure 3.7 Average Real-time Message Loss Rate vs. Total Traffic Load

Chapter 3 Providing Integrated Service in CDMA Wireless Networks

Figure 3.8 Average Message Delay vs. Total Traffic Load

Figure 3.9 Average non-real-time Message Delay vs. Total Traffic Load

As shown in Figure 3.8, the average message delay by the DQRA+AP/CDMA protocol with MUF scheduling is more than 40 time slots. This is more than two times of that by the original DQRAP/CDMA protocol and the DQRA+AP/CDMA protocol with MUF-FIFO scheduling. This extremely poor performance comes from the facts that the DQRA+AP/CDMA with MUF prefers the real-time traffic and blocks the non-real-time traffic transmission. This fact has been shown more clearly in the Figure 3.9, which demonstrates the relationship between the average message delay and the traffic load for

Chapter 3 Providing Integrated Service in CDMA Wireless Networks

the non-real-time traffic. The non-real-time messages have experienced the average delay of 70 time slots by the DQRA+AP/CDMA protocol with MUF scheduling. This is more than three times the average delay by the other two. A conclusion can be reached that the DQRA+AP/CDMA protocol with MUF scheduling is not suitable to support an integrated traffic due to its extreme preference.

Additionally, it is apparent that the DQRA+AP/CDMA protocol with MUF-FIFO scheduling has achieved a moderate improvement over the DQRAP/CDMA protocol in terms of more than 15% lower message loss rate while it has almost the same average message delay as that by the DQRAP/CDMA protocol even to non-real-time messages. It is proved a desired scheme that can achieve performance improvements for the real-time traffic over the DQRAP/CDMA protocol without much decrease on the performance of the non-real-time traffic in an integrated traffic environment.

Figure 3.10 Average Message Loss Rate vs. Traffic Load when Real-time Traffic Percentage is 20%

Figure 3.11 Average Message Loss Rate vs. Traffic Load when Real-time Traffic Percentage is 40%

Figure 3.12 Average Message Loss Rate vs. Traffic Load when Real-time Traffic Percentage is 60%

Figure 3.13 Average Message Loss Rate vs. Traffic Load when Real-time Traffic Percentage is 80%

Figure 3.14 Average Message Loss Rate vs. Traffic Load when Real-time Traffic Percentage is 100%

Figures 3.10 – 3.14 show the relationship between the average message loss rate and the traffic load for the DQRA+AP/CDMA protocol with MUF-FIFO scheduling under an integrated traffic environment with different percentages of real-time traffic. The message loss rate will increase in a reasonable amount correspondingly as the percentage of the real-time traffic increases. The loss rate will also increase when the total traffic load

Chapter 3 Providing Integrated Service in CDMA Wireless Networks

increases. However, the maximum loss rate is always under 12%, which is 15% lower than that by the DQRAP/CDMA protocol.

Figures 3.15 – 3.20 present the relationship between the average message delays and the traffic load for the DQRA+AP/CDMA protocol with MUF-FIFO scheduling under an integrated traffic with different percentages of real-time traffic. The average message delay will decrease as the percentage of the real-time traffic increases. It reveals that this algorithm prefers real-time traffic in nature. The average message delay by the algorithm is almost same or a little bit higher when the traffic load is heavy, and quite lower when the traffic is light, compared with that by the DQRAP/CDMA protocol, even when there is no real-time traffic. The beauty of the DQRA+AP/CDMA protocol with MUF-FIFO scheduling is that it has moderately reduced the average message loss rate without much increase on the average message delay. This feature could benefit non-real-time traffic in an integrated traffic environment while providing good service to real-time traffic at same time. It can be regarded as a potential candidate to support the integrated traffic with some extent of fairness in the CDMA wireless networks.

Figure 3.15 Average Message Delay vs. Traffic Load when Real-time Traffic Percentage is 0%

Figure 3.16 Average Message Delay vs. Traffic Load when Real-time Traffic Percentage is 20%

Figure 3.17 Average Message Delay vs. Traffic Load when Real-time Traffic Percentage is 40%

Figure 3.18 Average Message Delay vs. Traffic Load when Real-time Traffic Percentage is 60%

Figure 3.19 Average Message Delay vs. Traffic Load when Real-time Traffic Percentage is 80%

Figure 3.20 Average Message Delay vs. Traffic Load when Real-time Traffic Percentage is 100%

Figure 3.21 depicts the relationship between the message loss rate of the real-time traffic with the traffic load in the unit of number of messages per time slot under 60% of real-time traffic. It is clearly shown that the original DQRAP/CDMA protocol has not considered real-time service resulting in the message loss rate being comparatively large. This protocol will not favor the real-time traffic. However, the DQRA+AP/CDMA protocol with three scheduling algorithms have considered supporting real-time service in different ways. The results shown in the figure reveal that the performances of the DQRA+AP/CDMA protocol with three scheduling algorithms including the proposed two solutions in this paper have better performance for the real-time service. And both of them have almost behaved same as that of the DQRA+AP/CDMA protocol with MUF algorithm.

Figure 3.21 Average Real-time Message Loss Rate vs. Traffic Load with 60% Real-time Traffic

Figure 3.22 shows the relationship between average message delay of non-real-time traffic with the traffic load under 60% real-time traffic. It is observed that the average message delay for the DQRAP/CDMA protocol is higher than others when the traffic load is light. It is due to the high possibility of multiple collisions of the transmission requests. However, by the DQRA+AP/CDMA protocol, the possibility of the multiple collisions is almost zero so that performance of the DQRA+AP/CDMA protocol is better when the traffic load is light. It has shown that the performance in terms of the average message delay of the DQRA+AP/CDMA protocol with MUF scheduling is not as good as that of the original DQRAP/CDMA protocol, and even much worse, when the traffic is heavier. The MUF scheduling has only preferred to the real-time traffic so that it cannot benefit the non-real-time messages. On the contrary, the DQRA+AP/CDMA protocol with MUFC and CBPA scheduling can be expected to have the ability to serve the integrated traffic with reasonable performance due to their considerations on other parameters in scheduling. It is shown that the DQRA+AP/CDMA protocol with MUFC scheduling can

Chapter 3 Providing Integrated Service in CDMA Wireless Networks

relieve the increase of the delay. And the CBPA scheduling can keep much lower average message delay to the non-real-time traffic when the traffic load increases. It outperforms much over other scheduling algorithms. The better performances come from the functions of the MUFC and CBPA algorithms, which have provided transmission opportunities to the non-real-time messages while keeping the performance of real-time service unchanged. The performance of the CBPA algorithm is the best since it has more accurate consideration to schedule the transmission of real-time or non-real-time messages.

Figure 3.22 Average Non-real-time Message Delay vs. Traffic Load with 60% Real-time Traffic

Next, we present the performance of the two protocols with three scheduling algorithms under two different percentages of the real-time traffic. In Figure 3.23, we show the relationship between the message loss rate of the real-time traffic with the traffic load under 80% of real-time traffic. It is noticed that with increasing amount of real-time traffic, the real-time performance of the DQRA+AP/CDMA protocol with scheduling keeps always better than that of the DQRAP/CDMA protocol. And the CBPA algorithm has started to show that it is the best among the three scheduling algorithms on real-time

Chapter 3 Providing Integrated Service in CDMA Wireless Networks

performance. This fact comes from the reason that DQRA+AP/CDMA protocol with scheduling prefers real-time traffic. Since the majority of the network traffic load is the real-time traffic, the scheduling algorithms will benefit the real-time traffic much to show their effectiveness.

Figure 3.23 Average Real-time Message Loss Rate vs. Traffic Load with 80% Real-time Traffic

Figure 3.24 shows the relationship between average message delay of non-real-time traffic with the traffic load under 80% real-time traffic. The performance of both protocols and all three algorithms have kept the same trends as those when the percentage of the real-time traffic being 60% of the entire traffic load. But the performances of all the DQRA+AP/CDMA protocol with scheduling algorithms have been improved anyway. It is observed that although the DQRA+AP/CDMA protocol with MUF scheduling still performs the worst to the non-real-time traffic, the performance of the scheduling has been improved much in the sense that the average message delay has been largely reduced when the amount of the real-time traffic becomes larger. It implies that the

Chapter 3 Providing Integrated Service in CDMA Wireless Networks

performance of the DQRA+AP/CDMA protocol with scheduling algorithms become much better than that when the percentage of the real-time traffic is smaller. It reveals that larger amount of the real-time traffic can make the DQRA+AP/CDMA protocol with scheduling algorithms more effective resulting in their better performances to both real-time traffic and non-real-time traffic.

Figure 3.24 Average Non-real-time Message Delay vs. Traffic Load with 80% Real-time Traffic

In Figure 3.25, we show the relationship between the message loss rate of the real-time traffic with the traffic load under 20% of real-time traffic. It is noticed that with decreasing amount of real-time traffic, the real-time performance of the DQRA+AP/CDMA protocol with scheduling will perform poorer than those when the real-time traffic occupies 60% or 80%. The message loss rate by all the scheduling algorithms with the DQRA+AP/CDMA protocol come close to the message loss rate by the DQRAP/CDMA protocol, which has shown to be the worst performance.

Figure 3.25 Average Real-time Message Loss Rate vs. Traffic Load with 20% Real-time Traffic

Figure 3.26 shows the relationship between average message delay of non-real-time traffic with the traffic load under 20% real-time traffic. The performance of both protocols and all three algorithms have kept the same trends as those when the percentage of the real-time traffic being 60% or 80% of the entire traffic load. However, the performances of all the DQRA+AP/CDMA protocol with scheduling algorithms have become poorer in the sense that the average message delays by all the scheduling algorithms with the DQRA+AP/CDMA protocol have got increased although the performance of the DQRA+AP/CDMA protocol with the CBPA scheduling is still the best. It has revealed the fact that once the majority of the entire traffic load is non-real-time traffic, the DQRA+AP/CDMA protocol with scheduling will not have much effect to the improvement on the performance of both real-time and non-real-time services.

Figure 3.26 Average Non-real-time Message Delay vs. Traffic Load with 20% Real-time Traffic

Figures 3.27 – 3.31 show the relationship between average message delays of total traffic with the traffic load by the CBPA algorithm when the percentage of the real-time traffic changes. Because the real-time messages will be dropped if they cannot start transmission before deadline, the effective real-time traffic load is smaller than the actual arrival. On the contrary, there is no discarding for non-real-time traffic so that all non-real-time messages have to be delivered no matter how long they are delayed in the system. We can verify these when the total arrival rate is 16 packets/slot that equals to the maximum service capacity. When the percentage of the real-time traffic is 100%, the average delay of all transmitted messages is around 12 time slots. At this time, the effective throughput is smaller than 16 packets/slot. When the percentage of the real-time traffic is 0%, the maximum of average message delay is infinite that cannot be displayed in the figure. The system is unstable because the effective traffic intensity equals to one at that time. Other

Chapter 3 Providing Integrated Service in CDMA Wireless Networks

values of the average delay increase as the percentage of real-time traffic increases and are bounded in the area between two conditions.

Figure 3.27 Average Real-time Message Delay vs. Traffic Load when Real-time traffic percentage is 20%

Figure 3.28 Average Real-time Message Delay vs. Traffic Load when Real-time traffic percentage is 40%

Figure 3.29 Average Real-time Message Delay vs. Traffic Load when Real-time traffic percentage is 60%

Figure 3.30 Average Real-time Message Delay vs. Traffic Load when Real-time traffic percentage is 80%

Chapter 3 Providing Integrated Service in CDMA Wireless Networks

Figure 3.31 Average Real-time Message Delay vs. Traffic Load when Real-time traffic percentage is 100%

Chapter 4 Providing Differentiated Services in CDMA

Wireless Networks

4.1 Introduction

The differentiated service (DiffServ) architecture is one of the two quality of service architectures in the Internet context that have been proposed by the IETF. It is based on a model where traffic is classified and conditioned at the edges when entering a DS domain, then marked and assigned to different behavior aggregates by setting the value of the DiffServ Code Point (DSCP). Within the core of the network, messages are forwarded according to the per-hop behavior (PHB) associated with the DSCP.

Two approaches for differentiated service have been defined: *absolute differentiated services* and *relative differentiated services*. The first one relies on admission control and resource reservation mechanisms in order to provide guarantees and statistical assurances for performance measures, such as minimum service rate and end-to-end delay. The approach of relative differentiated services groups network traffic in a small number of classes. These classes are ordered based on their message forwarding qualities, in terms of queuing delays and message losses. This classification should guarantee that high-priority classes ($j > i, \forall i, j \in \{1, \dots, R\}$, where R denotes the number of traffic classes) will offer better QoS than low-priority classes that meets their requirements, cost, and policy constraints. From this point of view, the relative service differentiation model follows the

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

architectural principle of *end-system adaptation*, since it provides the class selection as an additional dimension in the end-system adaptation space.

The basic premise of effective relative differentiated services architecture has to be:

- *Predictable*, in the sense that the differentiation should be consistent (i.e., higher-priority classes are better, or at least no worse) independent of the class load variations;
- *Controllable*, meaning that the network operators should be able to adjust the relative QoS difference among classes based on their criteria.

Dovrolis *et al* have presented the *proportional differentiation model* [24]. According to this model, the basic performance measures for message forwarding locally at each hop are ratioed proportionally to certain *class differentiation parameters* that the network operator chooses. It is generally agreed that a better network service means lower queuing delays and lower likelihood of message losses. Dovrolis *et al* propose and study the proportional delay differentiation (PDD) model. The PDD model controls the ratios of the average queuing delays among classes, based on the specified delay differentiation parameters (DDPs). The proportional average delays (PAD) scheduler appears to always meet the PDD model. But it exhibits unpredictable behavior in short timescales.

In our hierarchical scheduling design for integrated traffic composed of real-time and non-real-time messages in CDMA wireless networks, the PDD model is considered as a desirable inter-class scheduling algorithm for multiplexing messages from the different classes to the server so that the system can support the differentiated service among classes with different QoS requirements. Different methods are proposed to control the

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

service order of the messages within the same class to improve the performance of that class in terms of average waiting delay. These intra-class scheduling algorithms are to guarantee minimum end-to-end delay and deadline violation probability. We combine the intra-class and inter-class scheduling methods to propose a novel MAC protocol for wireless networks that benefit real-time messages without violating the QoS requirement of non-real-time messages.

We also introduce a connection admission control (CAC) algorithm to control the traffic from different mobile users (MS). Based on the study of the proposed scheduling algorithm we establish a relationship between the message delay and traffic load. The CAC algorithm will reject the traffic from the MS whose QoS requirement cannot be supported. The objective of CAC is to accommodate as much traffic as possible while meeting all their QoS requirements at the same time.

The structure of this chapter is as follows. In Section 4.2, we present the PDD model and the PAD method to implement the model. In Section 4.3, we propose the novel MAC protocol for CDMA wireless networks. Section 4.4 presents the proposed CAC algorithm for the protocol. Performance evaluation and conclusion will be in Section 4.5.

4.2 Proportional Delay Differentiation

4.2.1 PDD model

We consider the following queuing model of a message multiplexer. A *no-drop*, *work-conserving*, and *non-preemptive* link with capacity C (bytes per second) serves N first-come-first-served (FCFS) queues, one for each traffic class. The *message scheduler* S

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

determines the order in which messages are selected for service. The no-drop property requires that the traffic intensity is less than 100% and enough buffers to queue message before transmission. The work-conserving property determines that the scheduler is not idle when there are messages in queue for service. The non-preemptive property means that a message transmission is always carried out to completion.

The interarrivals between messages of class i traffic flow are independent and identical distribution (IID) random variables, generated by a renewal arrival process for that class. The message sizes in class i are also IID random variables, generated by a renewal message size process for that class. Generally, the arrival and the message size processes of different classes are independent and different. The *aggregate traffic stream* that arrives in the queuing system is determined by the superposition of the N message streams that arrive in each class. Let λ_i be the average arrival rate in class i (messages per second), and $\lambda = \sum_{i=1}^N \lambda_i$ the average rate of the aggregate traffic stream. Let L_i be the average size (bytes) of class i messages, and $\bar{L} = \sum_{i=1}^N (\lambda_i \bar{L}_i) / \lambda$ the average message size among all classes messages. The link *utilization* is $u = \bar{L}\lambda / C < 1$.

The PDD model aims to control the *ratios* of the average class queuing delays based on the DDPs $\{\delta_i, i = 1, \dots, N\}$. Specifically, let $\bar{d}_i$ be the *average queuing delay*, or simply average delay, of the class i traffic. The PDD model requires that the ratio of average delays between two classes i and j is fixed to the ratio of the corresponding DDPs

$$\frac{\bar{d}_i}{\bar{d}_j} = \frac{\delta_i}{\delta_j} \quad i \leq i, j \leq N. \quad (4-1)$$

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

The DDP represents the priority of one class. Our convention is that higher-priority classes provide better service, i.e., lower queuing delays, and so $\delta_1 > \delta_2 > \dots > \delta_N > 0$. In the following, we choose Class 1 as the ‘reference class’ and set $\delta_1 = 1$. Then, the PDD model requires that the average delay of class i is a certain fraction δ_i of the average delay of Class 1

$$\bar{d}_i = \delta_i \bar{d}_1 \quad (4-2)$$

independent of the aggregate traffic load.

4.2.2 Performance of the PDD Model

When the traffic load of each class is defined, the $N-1$ ratios of the PDD model specify uniquely the average delays of the N classes. The key additional relationship in the mapping from delay ratios of class delays is the *conservation law* [25], [26], which constrains the average class delays in any work-conserving scheduler. Specifically, the conservation law requires that for any work-conserving scheduler S that causes an average delay $\bar{d}_i$ in class i , we must have that

$$\sum_{i=1}^N \lambda_i \bar{L}_i \bar{d}_i = \lambda \bar{L} \bar{d}_{ag} = \bar{q}_{ag} \quad (4-3)$$

where $\bar{d}_{ag}$ and $\bar{q}_{ag}$ is respectively the average delay (in seconds) and the average backlog (in bytes) in a FCFS scheduler that has the same capacity and serves the same aggregate traffic stream as S . The conservation law (in the pervious form) holds under arbitrary distributions for the message interarrivals and message sizes, as long as the first moment of these distributions (λ_i and $\bar{L}_i$) exist. The conservation law implies that even though a

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

scheduler S can affect the relative magnitude of the class delays, making the delay of one class lower than the delay of another class, this is a ‘zero-sum’ game because the weighted sum of (4-3) has to be equal to the average backlog $\bar{q}_{ag}$ of the aggregated traffic stream. The average backlog $\bar{q}_{ag}$ is a significant *invariant* in this balance, as it does not dependent on the scheduling discipline, but only on the aggregate traffic stream and on the capacity C .

Suppose now that the scheduler S satisfies the PDD model of (4-1). With the additional constraint (4-3) that the conservation law impose, we can show that the average delay in class i is

$$\bar{d}_i = \frac{\delta_i}{\sum_{n=1}^N \delta_n \lambda_n \bar{L}_n} \quad i = 1, \dots, N. \quad (4-4)$$

Consequently, even though the PDD model consists of $N-1$ *relative* constraints, the average delay in each class can be determined when the PDD model is applied to a certain traffic stream with class arrival rate $\{\lambda_i\}$, average message sizes $\{\bar{L}_i\}$, and average aggregate backlog $\bar{q}_{ag}$.

4.2.3 Features of the PDD Model

In this section, we further assume that the N classes have the same message size distribution. We then set $\bar{L}_i = \bar{L} = 1$, and normalize the backlog measures to *average message* units. The average delay in each class is then simplified to

$$\bar{d}_i = \frac{\delta_i}{\sum_{n=1}^N \delta_n \lambda_n} \quad i = 1, \dots, N. \quad (4-5)$$

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

Additionally, we assume that the N classes have the same interarrival distribution, but with a possibly different average interarrival $1/\lambda_i$.

Based on (4-5), we can now investigate the variations in the average class delay under the constraints of the PDD model as the DDPs vary. The following properties are referred as the *delay dynamics in the PDD model*.

Property 1: The average delay of a class increases as the arrival rate of this class increases. This property shows that there is no isolation between classes in the PDD model, i.e., when the delay of a class increases, due to additional load in that class, the average delays of all classes will also increase.

Property 2: Increasing the arrival rate of a higher-priority class causes a larger increase in the average class delays than increasing the arrival rate of a lower-priority class. In the case of two classes, for instance, suppose that the class arrival rates become either $\lambda'_1 = \lambda_1 + \varepsilon$ and $\lambda'_2 = \lambda_2$, or $\lambda''_1 = \lambda_1$ and $\lambda''_2 = \lambda_2 + \varepsilon$ ($\varepsilon > 0$). Even though the conservation law requires that the weighted average of the class delay is the same in both cases ($\lambda'_1 \bar{d}'_1 + \lambda'_2 \bar{d}'_2 = \lambda''_1 \bar{d}''_1 + \lambda''_2 \bar{d}''_2$), the class average delays in the second case are larger, i.e., $\bar{d}''_1 > \bar{d}'_1$ and $\bar{d}''_2 > \bar{d}'_2$, because of the PDD constraints. This property presents that higher-priority classes cost more, in terms of queuing delay, than lower-priority classes.

Property 3: Decreasing the delay differentiation parameter of a class will decrease the average delay of that class, and increase the average delay of all other classes. This property implies that if the delay of a class is reduced by lowering its DDP, the delay of all other classes will increase.

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

We have assumed so far that the PDD model is *feasible*, i.e., that there exists a work-conserving scheduler that can meet the constraints of (4-2). When the traffic percentage and the average backlog of the aggregate traffic are determined, the PDD model specifies the corresponding average delays of N classes as (4-4). However, there may not exist a work-conserving scheduler that can set the average delay of each class as in (4-4). Intuitively, the reason is that the average delay of a class has a minimum value due to the inherent load in that class. This minimum average delay available would be result when that class was given strict priority over the rest of the traffic.

Formally, given the arrival rates $\{\lambda_i\}$, the average class messages sizes $\{\bar{L}_i\}$, and the average backlog $\bar{q}_{ag}$ of the aggregate traffic stream, we say that a set of DDPs $\{\delta_i, i = 2, \dots, N\}$ is feasible when there exists a work-conserving scheduler that can set the average delay of each class as in (4-4). So, the set of DDPs $\{\delta_i, i = 2, \dots, N\}$ is feasible when the set of class delays $\{\bar{d}_i = \delta_i \bar{q}_{ag} / (\sum_{n=1}^N \lambda_n \delta_n \bar{L}_n), i = 1, \dots, N\}$ is feasible.

The necessary and sufficient conditions for the feasibility of a set of N average class delays, given the N the class traffic loads, were derived by Coffman and Mitrani. Under general assumptions that the interarrival and message size distributions are both Poisson distributions, a set of N average delays $\{\bar{d}_i, i = 1, \dots, N\}$ is feasible if and only if the following $2^N - 2$ inequalities hold:

$$\sum_{i \in \phi} \lambda_i \bar{d}_i \bar{L}_i > \bar{d}_\phi^{SP} \sum_{i \in \phi} \lambda_i \bar{L}_i = \bar{q}_\phi^{SP} \quad \text{for all } \phi \in \Phi, \quad (4-6)$$

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

where Φ is the set of $2^N - 2$ nonempty proper subsets of the set $\{1, 2, \dots, N\}$. $\bar{q}_\phi^{SP}$ and $\bar{d}_\phi^{SP}$ are the average backlog and the average delay, respectively, of the traffic in the set ϕ , if ϕ was given strict priority over all other classes. The Coffman-Mitrani inequalities state that any set of classes ϕ has a lower bound on its average backlog and this minimum value results when the traffic in ϕ is served with the highest priority in a Strict Priority (SP) scheduler.

4.3 Hierarchical Scheduling for Integrated Traffic in CDMA Wireless Networks

In this section, we propose a hierarchical scheduling algorithm in the CDMA wireless networks. The whole network environment is same as the DQRA+AP/CDMA protocol in Chapter 3. There are scheduling issues involved in both the collision resolution and the data transmission procedures. We keep the request resolution algorithm in the DQRA+AP/CDMA protocol and introduce the hierarchical scheduling algorithm in the data transmission procedure so that the novel MAC protocol can provide the differentiated service in the CDMA wireless networks.

4.3.1 System Model and Service

It is assumed that there are two kinds of traffic, real-time and non-real-time messages, integrated in the network. Each real-time message has its time constraint represented by laxity, which is time duration from the message enters the network to the time this message should begin transmission, while the non-real-time messages have no such delay requirements.

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

There are two distributed queues residing at each mobile node, the collision resolution queue and the data transmission queue, as the DQRA+AP/CDMA protocol. Each node has four variables to indicate the network situation regarding the two distributed queues. They are TQ , RQ , pTQ , and pRQ . The meaning and the updating rules of these variables are same as DQRA+AP/CDMA. In addition, the network architecture can further be modeled as a queuing system that consists of two subsystems: the collision resolution subsystem and the transmission subsystem. The system model of the RQ and TQ can be expressed in Figure 3.1.

When a message is generated at a mobile node, the node will first randomly select a spreading code as a transmission channel and send a request to indicate its data transmission in one of the control minislots pertaining to this channel. The transmission of the requests in the control minislots will be detected by the base station. Then the base station will broadcast the feedback information on the states of the control minislots to mobile nodes in the downlink channel, before next time slot begins. If it is indicated that a request collides with one or more requests from other mobile nodes, all the collided nodes will enter the collision queue in a certain order waiting for collision resolution. On the contrary, if a request has been successfully transmitted, the corresponding node will enter the data transmission queue and wait to send data to the base station.

Once a collided node has entered the collision resolution queue, it will not experience a competition for a control minislot any more. Rather it will be directly assigned into a control minislot in the next or following time slots. Collision resolution and data transmission processes work in parallel. Each message in the data transmission queue will

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

be transmitted through any logical channel once enough data transmission slots for the message of a certain channel are available. One message will be broken into packets for transmission as the data transmission slot duration is defined as long as the transmission time of one packet. However, before the transmission, a message with time constraint will be checked to tell whether it has been delayed in the queues longer than its time constraint. If yes, the message will be dropped. Otherwise, it can be successfully transmitted to the base station. During the collision resolution and transmission period, only the mobile nodes, which are not in either the collision resolution queue or the data transmission queue, can send requests to the base station for data transmission so that the total amount of traffic could be controlled to make the network stable.

4.3.2 Scheduling Issues and Algorithm

The data transmission procedure has been divided into two hierarchical levels, inter-class and intra-class, when there exist multiple classes of traffic in the network. The inter-class schedulers multiplex the messages from different classes to the server. In other words, the inter-class level scheduling algorithms are supposed to provide the service differentiation according to different QoS requirements. The controller on this level focuses on the priorities among the various parallel classes without considering the service order of the messages within a specific class. On the other hand, the intra-class schedulers decide the service priorities of messages within the same class without the effect on the scheduling of different classes. The objective of the lower level scheduling is to provide required service order according to special needs of the specific class. We combine different inter-class scheduling methods to support the differentiation (or balance) among different classes with intra-class scheduling methods to provide required service for particular

class. The detailed structure of the combined scheduling algorithms is shown as following figure.

Figure 4.1 Structure of Scheduling Algorithm Combined of Inter- and Intra-Class Schedulers

Here the inter-class scheduler selects the messages for service from one class i ($1 \leq i \leq N$) as soon as the resource is available. And the intra-class i decides which waiting message in the queue i will be served first when the resource is available. So there are total $N+1$ schedulers if there exist N different classes in the network. The scheduling methods of different intra-class scheduler i and j ($1 \leq i, j \leq N$) may be same or different, which are decided by the particular requirements of the class i and j ($1 \leq i, j \leq N$).

We consider a simple condition in the later study. There are only two different traffic types in the system, which means $N=2$ in proposed combined scheduling algorithm. One traffic class is composed of soft real-time messages and the messages failed to transmit before deadline will not be discarded but still served. The other traffic type is composed of non-real-time messages. They are no time limitation and need a best-effort service.

4.3.2.1 Inter-Class Scheduling

With our objective to provide service differentiation between the real-time and non-real-time classes, the *proportional delay differentiation* (PDD) model is desirable to provide scalable, predictable, and controllable service architecture. One way to interpret the PDD model is that the *normalized average delays*, defined as $\tilde{d}_i = \bar{d}_i / \delta_i$, must be equal in all classes, i.e.,

$$\tilde{d}_i = \frac{\bar{d}_i}{\delta_i} = \frac{\bar{d}_j}{\delta_j} = \tilde{d}_j \quad 1 \leq i, j \leq N. \quad (4-6)$$

This scheduling algorithm is referred as proportional average delay (PAD) scheduling.

Let $B(t)$ be the set of backlogged classes at time t , $D_i(t)$ be the sequence of Class i messages that departed before time t in order of their departure, and d_i^m be the delay of the m th message in $D_i(t)$. Assuming that there is at least one departure from class i before t , the normalized average delay of class i at t is

$$\tilde{d}_i(t) = \frac{1}{\delta_i} \frac{\sum_{m=1}^{|D_i(t)|} d_i^m}{|D_i(t)|} = \frac{1}{\delta_i} \frac{S_i}{P_i} \quad (4-7)$$

where S_i is the sum of queuing delays of all messages in $D_i(t)$, and $P_i = |D_i(t)|$ is the number of messages in $D_i(t)$. Suppose that a message has to be selected for transmission at time t . PAD algorithm chooses the backlogged class j with the maximum normalized average delay,

$$j = \arg \max_{i \in B(t)} \tilde{d}_i(t). \quad (4-8)$$

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

The message at the head of queue j is transmitted, after a queuing delay $d_j^{P_j+1}$. The variables S_j and P_j are then updated as $S_j + d_j^{P_j+1}$ and $P_j = P_j + 1$, and the new normalized average $\tilde{d}_j(t)$ is computed from (4-7). The basic idea in PAD is that if a message from class j with the maximum normalized average delay is served, the delay of that message will not increase any more since it is served, and thus the increase of S_j due to that message will be minimized. So, serving a message from class j tends to reduce the difference of $\tilde{d}_j(t)$ from the normalized average delays of the other classes.

PAD, however, exhibits a pathological behavior in short timescales. Dovrolis et al study the simulation results and state that because PAD occasionally allows higher-priority classes to experience much larger delays than their long-term average delays because the scheduling is unaware of the waiting times of backlogged messages. It is also instructive that PAD does a poor job in meeting the PDD constraints in relatively short timescales. This anomalism should be expected because PAD attempts to equalize the *long-term* normalized average delays but not the normalized average delays in the *last K departures*. A good scheduler for the PDD model should not only achieve or closely approximate the constraints of (4-2), but it should also provide predictable delay differentiation in short timescales.

The inter-class scheduler is a delay-based differentiated service, and it provides a simple approach to relative DiffServ model. However, it focuses on the relative performance, or the ratio of average delay among the different classes, but not the absolute end-to-end delay guarantee. In general, we assume that an arriving message or customer is assigned a

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

set of parameters (either at random or based on some property of the customer) that determine his relative position in the queue through the decision rule known as the queuing discipline. This queuing discipline simply controls the operation of the corresponding queue without affecting other queues. As a result, we consider the queuing discipline as intra-class scheduling discipline if a specific class is regarded to be one queue. Clearly, the intra-class schedulers work in the level different from the inter-class schedulers. Therefore, it is possible to shorten the absolute delays of the classes at same time but the ratio between them still keeps unchanged. Also the efficient queuing algorithm may reduce the ratio of the real-time messages that cannot transmit timely before expiring the time constraints while not impact the service order of non-real-time messages in other queue. Thus, we may propose the specific intra-class scheduling schemes according to the different requirement of each class.

4.3.2.2 Intra-Class Scheduling

The design objective of the intra-class scheduler for real-time messages is to reduce the tardy rate while the average queuing delay will not increase much or be even shorter. On the other hand, the non-real-time messages have no real-time constraints. They are served in a “best-effort” way. Thus the goal of the scheduling scheme for non-real-time class is to shorten the average waiting time as much as possible. The average delays of real-time and non-real-time classes are dependent on each other because the ratio between them keeps unchanged. So changing the intra-class queuing algorithm for one class will lead to the change in the average queuing delay of the other class at same time. Then the great decrease in the average message delay of non-real-time traffic will result in the decrease in the average message delay of real-time traffic though the intra-class scheduler for non-

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

real-time class have no impact on the service order of the real-time queue, which is determined by the inter-class scheduler that strictly controls the proportional delay differentiation between the classes.

4.3.2.2.1 Real-Time Traffic Scheduling

In this part, we assume that the messages that already miss the deadline before they begin transmission will not be discarded, which is different from the presupposition in Chapter 3. The network will mark the messages that cannot be delivered timely as “unsuccessful”, but they will still be served as same as the “successful” messages. In other words, the real-time messages in our network are soft real-time messages. The tardy rate is an important performance measurement for soft real-time network similar as the loss rate for hard real-time messages. The rate is the ratio of the numbers of the messages later than their deadlines to total served messages as shown in the following equation,

$$R_{td}(t) = \frac{N_{un}(t)}{N_{total}(t)}, \quad (4-9)$$

where $R_{td}(t)$ is the tardy rate of real-time queue at time t , $N_{un}(t)$ is the number of the unsuccessful real-time messages till time t and $N_{total}(t)$ is the total number of the real-time messages served till time t .

To overcome the problem that the average delay increases sharply under the heavy load, we present another method to debase the impact resulted by MUF and it is different from the one combined of MUF and FCFS in Chapter 3. With original MUF discipline, a deadline is assigned to each message before which it must be transmitted. We assume that the k th real-time message has its own delay bound d^k . This bound is exponentially

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

distributed with the mean value $\bar{d}$. The message is stamped at the arrival instance with the following timestamp:

$$D^k = a^k + d^k \quad (4-10)$$

D^k is the k th message's deadline set to the beginning of the service. MUF scheduling transmits messages in increasing order of its deadline. MUF can also be seen as a Time-Dependent Priority (TDP) scheduler, where the priority of each message is given by:

$$P^k = D^k - t_{now}, \quad (4-11)$$

where t_{now} is the current time.

Kleinrock states in [25] that the average waiting time using FCFS is longer than that for Shortest-Processing-Time-First (SPTF). The relative position of the message in the queue is determined by its service time. If we suppose that the capacity of one server is unity, the normalized service time is equal to the message length.

$$P^k = l^k, \quad (4-12)$$

where P^k is the priority of the k th message, and l^k is the k th message's length and is exponentially distributed with mean value $\bar{l}$. SPTF allocates the highest priority to the message with the shortest length.

MUF is effective to reduce the tardy rate of the real-time messages and SPTF is to reduce the average delay of the messages. It is possible to propose a scheduling algorithm combined of SPTF and MUF. By this method we can get the both advantages of the two algorithms. The proposed scheduler calculates the priorities of the messages as following:

$$P_{RT}^k = a^k + d^k + l^k = D^k + l^k \quad (4-13)$$

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

The queuing scheduler transmits the messages in increasing order of P_{RT}^k . In fact, P_{RT}^k can be decomposed into two parts: the deadline set to the beginning of service, D^k ; the message length, or normalized service time, l^k . Suppose there are two messages A and B arriving in the system and waiting for transmission at same time. If their deadlines are same, or $D^A = D^B$, the service order now is determined by the message length. The one with shorter length will obtain a higher priority than the longer one. Then the combined scheduler is simply a SPTF scheduler. On the other hand, if the two messages have the same service time or, $l^A = l^B$, the service order is decided by the deadline. The message with a more urgent time constraint will be allocated a higher priority than the looser one. Thus the proposed queuing scheme is as same as MUF scheduler.

The SPTF scheduler always chooses the messages in the queue with a higher priority over one of the lower priority. Therefore the lower priority may suffer starvation. The starvation for longer messages is possible to happen as long as there is a steady supply of shorter messages in SPTF. MUF scheduler will not lead to this situation because it can be considered as time-dependent priority scheduler. It allows one message to change the priority based on the message's age. Actually P_{RT}^k can be seen as the deadline set to the end of the service, however, the proposed algorithm is different from the conventional MUF scheduling. The original MUF scheduler that sets the deadline to the end of service will abandon the message as soon as it expires the deadline; even the transmission is in process. In our system the message will not be interrupted after the service begins as long as it does not miss the deadline at the time it starts transmission. From this point the proposed intra-class scheduler can be regarded as one variant of MUF so that it will not result in starvation problem.

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

In fact, this MUF+SPTF scheduling algorithm is an extension of the Cost-based Priority Assignment (CBPA) proposed in Chapter 3. CBPA determines the real-time message service priorities by the calculation of the normalized laxity and message length, as

$$p_i = \alpha \frac{\text{laxity}_i}{\text{laxity}_{\max}} + \beta \frac{\text{length}_i}{\text{length}_{\max}}. \quad \alpha \text{ and } \beta \text{ are the coefficients for the laxity and message}$$

transmission time. Different sets of α and β make the CBPA flexible and scalable. On

the other hand, in MUF+SPTF scheduling, real-time message's priority is defined

$$\text{as } p_i = \text{arrival time}_i + \text{laxity}_i + \text{length}_i. \quad \text{For CBPA, when } \frac{\alpha}{\beta} = \frac{\text{laxity}_{\max}}{\text{length}_{\max}},$$

$$p_i = \frac{\alpha}{\text{laxity}_{\max}} (\text{laxity}_i + \text{length}_i) \triangleq \text{laxity}_i + \text{length}_i \quad \text{since } \alpha \text{ and } \text{laxity}_{\max} \text{ are both}$$

predefined constants. At this time, $p_i^{\text{MUF+SPTF}} = \text{arrival time}_i + p_i^{\text{CBPA}}$. It is clear that the

priority definition in MUF+SPTF is an absolute time point. As the time elapses, this point

can be reached finally. However, the priority definition in CBPA is a relative value, or a

period. When α , β , $\text{laxity}_{\max}$, and $\text{length}_{\max}$ are fixed, this relative value can be viewed

as a transformation of the expected waiting time in the network. Elapsing time cannot

guarantee the message with a determined priority will be served in an explicit time. In

other words, this type of priority definition may cause starvation problem. Furthermore,

CBPA can complex integrated traffic of real-time and non-real-time. For non-real-time, if

the value of laxity is fixed to $\text{laxity}_{\max}$, CBPA scheduling is able to serve two kinds of

traffic efficiently and fairly under proper α and β . But MUF+SPTF is only suitable for

real-time traffic scheduling.

4.3.2.2.2 Non-Real-Time Traffic Scheduling

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

The non-real-time service is time insensitive and “best effort”. If w^i is the time spent by the i th arriving message waiting in queue (not including the transmission time), M^i is the message’s transmission time, and T^i denotes the total time spent in the system by the message, we have

$$T^i = w^i + M^i. \quad (4-14)$$

The intra-class scheduler for non-real-time class is to reduce the average delay, or $E[T^i]$. In the two components of T^i , the transmission time or the service time is a predetermined parameter associated with the message when it was generated. Therefore, the only way to improve the delay performance is to reduce w^i , the waiting time before the transmission.

If we assume that the service capacity is unity, the message’s service time M^i is equal to its size l^i . Back to equation (4-14), we replace M^i by l^i and obtain the expression

$$T^i = w^i + l^i. \quad (4-15)$$

Then T_i can be regarded as the expected delay of the message if it is permitted for transmission immediately. Moreover, the required bandwidth c^i for the server to deliver the i th message can be estimated by:

$$c^i = \frac{l^i}{T^i} = \frac{l^i}{w^i + l^i}. \quad (4-16)$$

Liu and Wong present in [30] an algorithm that estimates the real-time messages by the ratio: size/(deadline-current time), and admits those with minimum estimated bandwidth at highest priorities. We introduce the idea that the transmission of the non-real-time messages is in increasing order of their estimated bandwidth. Therefore, the scheduler

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

allocates the highest priority to the message with the smallest c^i , which is named as Smallest Estimated Bandwidth First (SEBF) scheduling algorithm.

$$P_{NRT}^i = c^i = \frac{l^i}{w^i + l^i} = 1 / \frac{w^i + M^i}{M^i} = \frac{1}{\frac{w^i}{M^i} + 1}. \quad (4-17)$$

Thus we establish a direct relationship between w^i and M^i that can discover the characteristics of the queue. Then the intra-class scheduler for non-real-time type allocates the highest priority to the message with the minimum P_{NRT}^i .

In fact, the proposed intra-class scheduler is an extension of SPTF scheduler. Again, suppose there are two non-real-time messages, A and B , waiting in queue for service. If the two messages arrive at the same time, then $w^A = w^B$, the priority is decided by the message length. The one with a shorter length will be allocated a higher priority than the longer one as SPTF scheduler does. On the contrary, if the two messages have the equal length, $l^A = l^B$, the priority is dependent on the time already spent in the queue. The system transmits firstly the message that waits a longer time for service. At this time it is as same as a FIFO scheduler. So, it can be seen as a combination of SPTF and FIFO. In a word, the scheduler prefers the message with the smallest length (service time) and the longest waiting time for transmission.

In addition, the proposed scheduler will not encounter the starvation problem as SPTF. The reason is that the messages' priorities will change according to their ages in the system. Under the condition that the waiting times are same, the shorter message is delivered first and the longer one is postponed for the next chance. However, the priority

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

increases as the delay increases. Then the impact of the service time on the priority is counteracted by the waiting time.

Although introducing FIFO to SPTF can avoid the starvation problem, it also results in another trouble for the network controller. Every time the scheduler determines the priorities for all the members in queue, it must first compute the normalized waiting time of each message and compare the results to decide the messages' relative position for the transmission. If there are n messages waiting in the queue, the scheduler have to complete n division actions and $\frac{n(n+1)}{2}$ comparison actions (when using bubble sorting algorithm)

before it can exactly determine all the messages' priorities. To finish the operations in a short time is not a trouble for network controller when n is small. On the contrary, if n is very big under a heavy traffic load, completing such many operations rapidly is burdensome for the controller.

We can decrease the frequency of refreshing the messages' priorities to reduce the controller's overhead. The scheduler does not trigger the scheduling every moment but in a longer interval. The interval can be a fixed long period that is decuple to one system time slot or even longer. In addition, the interval can be defined dynamically by some special events. In our proposed queuing discipline, the scheduler triggers the priority decision in the latter way.

The service discipline with a gate in [31] is a proposal to resolve the problem, which decomposes the scheduling in two stages as in Figure 4.2. We define two queues of a tandem for the messages waiting for service, a waiting buffer and a transmission queue.

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

The messages from the mobile terminals are first gathered in the unordered buffer before they are put in SEBF order in the transmission queue. The scheduler will choose the head message in the transmission queue for service as soon as resource available. The network controller empty the buffer and pushes all the messages in the buffer to the transmission queue in increasing order of P_{NRT}^i when there is no message in the transmission queue. This time point is the trigger time point for the messages scheduling described above. The interval between two scheduling time points is random depending on the transmission of messages in the last gate. However, it is not more than the one slot period so that this Gated Smallest Estimated Bandwidth First (GSEBF) algorithm lowers the frequency of the scheduling than the original SEBF algorithm.

Figure 4.2 A Queue with a Waiting Buffer and a Transmission Queue

4.3.2.3 The Integrated Scheduling Scheme

The protocol is proposed on the base of the DQRAP/CDMA protocol. Again it can be modeled as shown in Figure 4.3. It has two components: the collision resolution subsystem and the transmission subsystem. The former portion is the same as that adopted in DQRA+AP/CDMA already presented in Chapter 3. And the part of data transmission subsystem is the integrated scheduler composed of the inter-class scheduler and the intra-class schedulers described above.

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

Figure 4.3 System Model of the Proposed Protocol

In the DQRAP/CDMA protocol, the proposed scheme uses two concatenated distributed queues: the collision resolution queue and the data transmission queue. When a message arrives in the network, the corresponding terminal, following a certain set of rules described below, selects a spreading code and sends a request in one from the second to m th control minislots pertaining to this code. If it fails (i.e., the request collides with one or more requests from other messages), it enters the collision resolution queue. Collisions are then resolved in the fixed order by the queue discipline as in the DQRA+AP/CDMA protocol. However, the data transmission is different from that in DQRAP/CDMA. We split the data transmission queue into two parts as those in DQRA+AP/CDMA because there are two types of traffic in the system for differentiated service. One is for real-time traffic and the other is for non-real-time traffic. The preal-time traffic is just kept in one queue. However, the non-real-time traffic is kept in two concatenated subqueues, which is decided by the particular intra-class schedulers for each class. The first subqueue is a

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

temporary buffer in which the messages are queued in FIFO order followed by the second one in which the messages are in increasing order of P_{NRT}^i defined in (4-17). This second subqueue is the actual data transmission queue for non-real-time messages. If there is a free channel, the modified PAD scheduler selects a message from the head of the two data transmission subqueues according to $\frac{S_{RT}}{S_{NRT}}$ where S_i is the sum of delay of the served messages in class i . If $\frac{S_{RT}}{S_{NRT}} > \frac{\delta_{RT}}{\delta_{NRT}}$, the inter-class scheduler selects the head message in real-time data transmission queue for service; otherwise, the head message in non-real-time data transmission queue is chosen for delivery.

Like DQRAP/CDMA, all terminals have two integer counters, pRQ and RQ , to represent the collision resolution queue. RQ is the number of collisions in the distributed transmission queue and pRQ is the position of a given terminal in the collision resolution queue. However, there should be two counters for the two data transmission subqueues, TQ_{RT} and TQ_{NRT} . Also the position of the corresponding terminal in the data transmission queue, pTQ , has different meanings in the different cases. If the terminal is to transmit a real-time message, pTQ is its position in the real-time data transmission subqueue; on the contrary, pTQ is the terminal's position in the non-real-time data transmission subqueue if the terminal is to deliver a non-real-time message. These values range from 0, meaning that the terminal does not occupy any position in the corresponding queue, to TQ_i or RQ (respectively), 1 being the head position of the queue. RQ have the same value for all the terminals in the system (i.e., they represent *distributed* queues), while TQ_{RT} is same for all real-time terminals and TQ_{NRT} is same for all non-real-time terminals and the two TQ s

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

may be different from each other. pTQ and pRQ have a specific value for each terminal. These values are initially set to zero and must be kept updated using the feedback information sent by the base station, each slot, using a broadcast channel, and following a set of rules described below. It has to include a final-packet-bit for each node, which is the mark that all the data terminals must send when they are transmitting the last packet from one message. This flag bit must be ON in the last packet of each message, and must be OFF in all other packets.

Also the proposed algorithm consists of three sets of rules that each terminal has to follow at the end of each slot. However, in a different order of execution from the original DQRAP/CDMA, they are the queuing discipline rules (QDR), the request transmission rules (RTR), and the data transmission rules (DTR).

1) QDR

- a) By one unit for each control minislots in the success state, each station increments the value of TQ_{RT} if the terminal is to transmit a real-time message; otherwise, TQ_{NRT} is increased by one, taking into the account the feedback information from all the control minislots from one of the K spreading codes.
- b) For each message correctly received by the base station with the final-packet-bit set to ON, each station reduces the value of TQ_{RT} or TQ_{NRT} according to the different cases.
- c) If $RQ > 0$, each station decrements the value of RQ by $\min(RQ, K)$ units.

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

- d) Each station increments the value of RQ by one unit for each control minislot in the collision state, taking into account all the control minislots from any of the K spreading codes.
- e) Depending on its state, and the results of the control minislots, each station calculates the values for pTQ and pRQ . That is, if it has sent a request and this request has succeeded, it calculates its position among all the succeeding minislots and sets pTQ to the corresponding value at the end of TQ_{RT} (for real-time terminal) or TQ_{NRT} (for non-real-time terminal). On the other hand, if the request has collided, the terminal calculates its position among all the present collisions and sets pRQ to the corresponding value at the end of RQ . If it has not sent any request, then pTQ and pRQ follow the same update rules as TQ and RQ , respectively, but only if the initial values are other than zero.

2) RTR

- a) If $RQ < K$ and more than one terminals generate newly arriving message during the last time slot randomly selects one from the second to m th minislots of all K spreading codes (total $(m-1) \times K$ minislots available for competition) and transmits a request on it.
- b) Those terminals in the first $\min(RQ, K)$ positions of the collision resolution queue send the messages to the data transmission queues. If the terminal is to send a real-time message, its pTQ is calculated as P_{RT}^k defined in (4-13) and TQ_{RT} adds one unit; otherwise, the value of TQ_{RT} keep unchanged because at this time the terminal is in the temporary buffer in the non-real-time data transmission queue. Every time all the jobs in the actual non-real-time data transmission queue are

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

completed. The temporary buffer is emptied and all jobs in them are transferred to the actual non-real-time data transmission queue, and the values of pTQ are calculated as P_{NRT}^i defined in (4-17).

3) DTR

- a) If $TQ_{RT} + TQ_{NRT} < K$ and only one terminal generates newly arriving messages in the last time slot, this terminal directly selects the available channel to start transmission without sending request for resource competition. This keeps the free access feature of the original DQRAP/CDMA that allows the newly arrived messages to be transmitted immediately when traffic load is light. But it does not cause a collision in the data part of a slot as DQRAP/CDMA.
- b) Once a channel is available, the inter-class scheduler selects the first message in real-time data transmission queue for service if $\frac{S_{RT}}{S_{NRT}} > \frac{\delta_{RT}}{\delta_{NRT}}$; otherwise, it selects the head message in non-real-time data transmission queue for service.

4.4 Admission Control Algorithm

4.4.1 Admission Control in Wireless CDMA Networks

In CDMA networks, the quality of ongoing connections will decline if cell interference is allowed to increase, due to the soft capacity [32]. Therefore, we need admission control to limit the amount of interference in the system. Admission control needs to check that admission of a new connection will not sacrifice the planned coverage area or QoS of the ongoing connections. In 3G networks, such as UMTS, admission control is located at the radio network controller, where the load information from several cells can be obtained. Because many applications may request asymmetrical bandwidths at the uplink and the

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

downlink, the admission control should estimate the load increase that the new connection will cause separately for the uplink and the downlink. Thus, the admission control decision is made independently for each direction, e.g., in WCDMA-FDD or cdma2000.

In FDMA/TDMA-based mobile networks, we have pre-specified the capacity per cell. But CDMA has no hard limit on the capacity, which makes admission control a more complex issue for soft capacity management. Generally, the admission control schemes for CDMA can be classified into the following groups:

- Signal-to-interference ratio (SIR)-based admission control: it is made on an individual basis comparing mobile's SIR to a given threshold value at the base station. It refers to the uplink direction.
- Load-based admission control: in this case, a new call is admitted if the total load will not exceed the threshold value after the admission.
- Power-based admission control: in the uplink, a threshold for maximum allowed interference is predefined. A new user is admitted if the new total interference is lower than the threshold. In the downlink, it uses a similar approach for the admission control. The new user is admitted if the new total transmission power from the base station will not exceed the predefined threshold.

According to this classification, our proposed admission control algorithm is a load-based admission control in uplink direction. The QoS consideration is real-time message delay that is related to the total traffic load in the network. The controller limits the traffic load from all mobile terminals in such a level that the corresponding average real-time message delay is lower than the requirements of the traffic flows.

4.4.2 Proposed Admission Control Algorithm

In our admission control design, the major considerations are real-time traffic service commitments and network utilization. There is a tradeoff between them. And we should make a balance so that it can support as much real-time traffic transmissions within the requested time limitations as possible while the network utilization is at a high level.

We assume that each end user or terminal in the network is permitted to generate only one type of traffic flow, either real-time or non-real-time traffic. The terminal generates the real-time traffic flow is called a “real-time terminal”; on the other hand, the terminal is named “non-real-time terminal” if the flow from it is a non-real-time one. The traffic flows from real-time terminals have particular time requirements, each of which is called delay requirement. The real-time messages from the same terminal i have their laxities exponentially distributed with the mean value d_i . And the different real-time terminals may have different d_i . These real-time messages are required to start their transmission before they miss their deadlines. The value of deadline is the summation of the message arrival time to the network and its particular laxity requirement. If the message can start transmission before its deadline, we denote it “successful”. Otherwise, it is marked “failed” to meet the deadline but still be delivered after deadline. The presented scheduler is composed of the inter-class scheduler for class delay differentiation and the intra-class schedulers for each class special consideration. We can obtain a delay performance $\bar{D}_{RT}$ that is the average delay for the real-time class of traffic on the terminals.

One purpose of the admission control is to make the messages in the admitted traffic flow be transmitted “successfully” as much as possible. In other words, it is to pursue the

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

average tardy rate for real-time traffic as small as possible. It demands that d_i of all the admitted real-time traffic flows are not smaller than the estimated $\bar{D}_{RT}$, or

$$\min(d_i) \geq \bar{D}_{RT}. \quad (4-18)$$

where $\min(d_i)$ is the minimum value among all admitted flows' d_i s. It is not a strictly sufficient to assure every message can meet its time requirement, but it can reduce the expiration in a great degree. For a flow with $d_i < \bar{D}_{RT}$, the server cannot transmit most messages “successfully” in that flow so that it will lower the average tardy rate of all real-time traffic very much if admitted. Moreover, if admitted, the unexpected flows will share the link capacity with those are supposed to transmit successfully. They occupy the bandwidth that are in fact expected to allocate to the latter because the system will not drop the expired messages. Thus it postpones the transmission of the latter and increases the chance that the expected messages fail to start transmission before deadline. This is a serious resource waste which leads to the decrease of the tardy rate on the real-time messages.

We can estimate that $\bar{D}_{RT}$ varies in different network traffic loads. Also $\bar{D}_{RT}$ changes according to the variations of the real-time traffic percentage of the total traffic load. So it is supposed that there exist some relationship between $\bar{D}_{RT}$ and network traffic load and percentage of real-time traffic load. We suppose a function to indicate the relationship as follows

$$\bar{D}_{RT} = F(Tr_{total}, Per_{RT}), \quad (4-19)$$

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

where Tr_{total} is the total network traffic load, Per_{RT} is the percentage of all real-time traffic in the total traffic. Moreover, we can express Tr_{total} and Per_{RT} by the loads of real-time and non-real-time traffic as follows

$$Tr_{total} = Tr_{RT} + Tr_{NRT}, \quad (4-20)$$

$$Per_{RT} = \frac{Tr_{RT}}{Tr_{total}} = \frac{Tr_{RT}}{Tr_{RT} + Tr_{NRT}}, \quad (4-21)$$

where Tr_{RT} and Tr_{NRT} are traffic loads of real-time and non-real-time traffic respectively. Then Tr_{RT} and Tr_{NRT} can be calculated by the average arrival rates of real-time and non-real-time traffic,

$$Tr_{RT} = \frac{\lambda_{RT}}{\mu}, \quad (4-22)$$

$$Tr_{NRT} = \frac{\lambda_{NRT}}{\mu}, \quad (4-23)$$

where λ_{RT} and λ_{NRT} are respectively the average arrival rates of total real-time and non-real-time traffic, and μ is the average network service rate. Additionally, because the messages arriving in a Poisson process and their lengths are exponentially distributed, we calculate λ_{RT} and λ_{NRT} by the known Kleinrock Independence Approximation in which merging several message streams on a transmission line is suggested to have an effect akin to restoring the independence of the inter-arrival times and message lengths,

$$\lambda_{RT} = \sum_{i \in G_{RT}} \lambda_i, \quad (4-24)$$

$$\lambda_{NRT} = \sum_{j \in G_{NRT}} \lambda_j, \quad (4-25)$$

where G_{RT} denotes the group containing all the flows from real-time terminals while G_{NRT} denotes the group containing all the flows from non-real-time terminals.

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

λ_n ($n = i, j$) represents the average arrival rate of traffic flow n . Because the total network service capacity is fixed and the mean length of all the messages is predetermined and assumed to be same, Tr_{total} and Per_{RT} can further denoted by the average arrival rates of total traffic and real-time traffic,

$$Tr_{total} = \frac{\lambda_{total}}{\mu} = \frac{\lambda_{RT} + \lambda_{NRT}}{\mu}, \quad (4-26)$$

$$Per_{RT} = \frac{\lambda_{RT}}{\lambda_{RT} + \lambda_{NRT}}. \quad (4-27)$$

We substitute (4-26) and (4-27) into (4-19) and neglect the known μ to obtain

$$\bar{D}_{RT} = F\left(\lambda_{RT} + \lambda_{NRT}, \frac{\lambda_{RT}}{\lambda_{RT} + \lambda_{NRT}}\right) = f(\lambda_{RT}, \lambda_{NRT}). \quad (4-28)$$

Finally we establish a function to estimate $\bar{D}_{RT}$ with λ_{RT} and λ_{NRT} .

As described in the previous, $\bar{D}_{RT}$ is expected to be small so that more real-time flows can meet the condition defined in (4-18) and be admitted to the network. The purpose of the proposed scheduling algorithm is to support as much successful real-time transmission as possible. The proportional delay differentiation (PDD) can only be implemented under a heavy traffic load. Although Tr_{total} and Per_{RT} together decide $\bar{D}_{RT}$, we can observe that an increase in Tr_{total} results in a much bigger increase on $\bar{D}_{RT}$ than the effect that a increase in Per_{RT} inflicts on $\bar{D}_{RT}$. So the total network traffic load is dominant in determining $\bar{D}_{RT}$. Therefore, the system tends to reject a non-real-time flow but not a real-time one although they have the same average arrival rates (defined in (4-24)-(4-27) the same average arrival rate means the same traffic load in the network) so

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

that dropping of any one nearly leads to the same reduce on $\bar{D}_{RT}$. Suppose that there are two data flows, one is real-time traffic A and the other non-real-time traffic B , to select for rejection. They have the same average arrival rates so that $\lambda_A = \lambda_B$. If rejecting any one can reduce $\bar{D}_{RT}$ in the same degree, the proposed admission controller will select the non-real-time flow to discard as described above. However, if $\lambda_A < \lambda_B$ and rejecting any one both can meet the requirement of $\bar{D}_{RT}$. The controller faces a dilemma: if dropping the non-real-time flow, it can support as much real-time traffic as possible but the total traffic load is not the largest. On the contrary, it can achieve a higher utilization but the real-time traffic load is not most if discarding the real-time traffic. It is a trade-off between the network utilization and the real-time traffic load. We put the weight on the side of the network utilization in the design of admission control algorithm. In other words, we prefer high network utilization to high real-time traffic load. Therefore, the controller discards the real-time traffic flow A to reduce $\bar{D}_{RT}$ in the example. Finally, we also have to avoid an extreme condition that the real-time flow has to be served during a despitful time limitation and the network cannot support it whatever. At that time, it is similar as a deadlock condition and even removing all non-real-time traffic cannot resolve the problem. The only method to unlock is to discard the real-time flow with the most stringent delay requirement.

To achieve high network utilization is the guideline to select a particular flow to reject among all from the non-real-time terminals if necessary. As explained previously, high network utilization means high traffic load to be served while the traffic load is directly related to the average arrival rate. Then the selected flow must have the smaller average arrival rate than any others so that the remaining flows have the biggest aggregate arrival

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

rates. We name it as Smallest-Arrival-First-Drop (SAFD) discipline for non-real-time traffic. The condition is more complicated for real-time traffic, however, because the traffic load and the time constraints for real-time traffic should be both included in the consideration for admission. The ideal condition is that the average arrival rate and delay requirement of the chosen flow for rejection are at same time smaller than any others. Otherwise, a complex consideration is needed before rejection. Again suppose that there are two real-time flows for rejection, A and B , and their average arrival rates and laxities are λ_A , λ_B , d_A , and d_B respectively. λ_A and λ_B are the minimum two among all average arrival rates while d_A and d_B are the minimum two among all average laxities. We assume $\lambda_A > \lambda_B$ and $d_A < d_B$. We can estimate $\bar{D}_{RT}$ according to (4-28). It is supposed to obtain $\bar{D}'_{RT}$ if rejecting A and $\bar{D}''_{RT}$ if rejecting B . Because the variation of the network utilization dominates in the determination of $\bar{D}_{RT}$, we can anticipate that $\bar{D}''_{RT} < \bar{D}'_{RT}$ due to $\lambda_A > \lambda_B$. An extreme condition is $d_A < \bar{D}''_{RT} < \bar{D}'_{RT} < d_B$. If referring to (4-18), the controller has to discard A first and the requirement of $\bar{D}_{RT}$ can be met because $d_B > \bar{D}''_{RT}$. Referring to SAFD discipline, however, the controller should reject B first but it is useless to resolve the problem due to $d_A < \bar{D}'_{RT}$. Then the controller has to continue to reject A and the network utilization is lower than that in the condition that A is first rejected. We highlight the order of rejecting traffic flows because the rejected flow will not be reconnected in future if the network condition is stable. So each step should be very careful and any mistake will lead to an unstable network condition or a serious resource waste. To avoid the condition that a real-time flow with regular arrival rate but an extremely large delay requirement or with a regular delay requirement but an extremely small arrival rate is admitted to the network, we design a selection method named

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

Smallest-Estimated-Buffer-First-Drop (SEBFD). It makes the decision to reject real-time traffic referring to the average arrival rate and delay requirement synthetically. The average delay requirement of a flow can be regarded as its expected average waiting time in queue, and with Little's Theorem we have

$$Buff_i = N_i \times l_i = \lambda_i \times d_i \times l_i, \quad (4-29)$$

where $Buff_i$ is the size of the estimated buffer to restore the data of flow i , N_i is the average number of messages from i , and l_i is the mean length of the messages of i . If the messages of all flows are exponentially distributed with the same mean value, (4-29) can be further simplified and defined as follow

$$Buff_i \triangleq \lambda_i \times d_i, \quad (4-30)$$

Following is the description of the proposed admission control algorithm with pseudo-code

```

/*  $\lambda_{RT}$  is the total arrival rate of all the real-time flows  $\lambda_{RT} = \text{sum}(\lambda_i)$ 

 $\lambda_{NRT}$  is the total arrival rate of all the non-real-time flows  $\lambda_{NRT} = \text{sum}(\lambda_j)$ 

 $\bar{D}_{RT} = f(\lambda_{RT}, \lambda_{NRT})$  */

/*  $\min(d_i)$  is the minimum value among all real-time average delay requirement */

while  $\bar{D}_{RT} > \min(d_i)$ 
{
    if  $f(\lambda_{RT}, 0) > \min(d_i)$ 

        /* SEBFD discipline */

        remove the real-time flow with the minimum  $\lambda_i \times d_i$  from the network;

    else

```

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

```
{  
    search the real-time flow  $i$  with the minim  $\lambda_i \times d_i$ ;  
  
    if( $\min(\lambda_j) > \min(\lambda_i)$ )  
        /* SAFD discipline */  
        remove the non-real-time flow  $j$  from the network;  
    else  
        /* SEBFD discipline */  
        remove the real-time flow  $i$  from the network;  
}  
  
}
```

admit all remaining flows to the network for service;

Figure 4.3A Procedures of the Propose Admission Control Algorithm

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

In addition, the admission control algorithm should support another condition different from the above one. If there are users in the network who have already been served stably, the new users' requests for service must be checked carefully before admission. The most important rule is that the service of already admitted users should be protected against introducing the new users. If accepting the new users lowers the QoS of the admitted ones, these latter ones cannot be all admitted. In the extremely rigorous case, any service request from the new users cannot be accepted if any admission will impair the QoS of the admitted ones. Because the admission and rejection only happen in the pool of the new users, the process of admission control should be limited in this small pool. The process will not administrate those admitted users. At the same time, the QoS requirements of the newly admitted users should also be guaranteed after the admission. Therefore, the operations of the admission control algorithm should be modified correspondingly. Pseudocode for the admission control operations is changed as following,

```

/*  $\lambda_{RT}$  is the total arrival rate of all the admitted real-time flows  $\lambda_{RT} = \text{sum}(\lambda_i)$ 

    $\lambda_{NRT}$  is the total arrival rate of all the admitted non-real-time flows  $\lambda_{NRT} = \text{sum}(\lambda_j)$ 

*/

/*  $d_1$  is the minimum delay requirement among all the admitted real-time flows */
for all new requests for real-time service
    reject those with smaller delay requirement requirements than  $d_1$ 

/*  $\lambda'_{RT}$  is the total arrival rate of all the remaining new real-time flows  $\lambda'_{RT} = \text{sum}(\lambda'_i)$ 

    $\lambda'_{NRT}$  is the total arrival rate of all the new non-real-time flows  $\lambda'_{NRT} = \text{sum}(\lambda'_j)$ 

*/

```

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

```

while  $\bar{D}'_{RT} = f(\lambda_{RT} + \lambda'_{RT}, \lambda_{NRT} + \lambda'_{NRT}) > d_1$ 
{
    search the real-time flow  $i$  with the minim  $\lambda_i \times d_i$  in the new arrived real-time flows;
    search the non-real-time flow  $j$  with the minim  $\lambda_j$  in the new arrived non-real-time
    flows;
    if( $\min(\lambda_j) > \min(\lambda_i)$ )
        /* SAFD discipline */
        remove the non-real-time flow  $j$  from the network;
    else
        /* SEBFD discipline */
        remove the real-time flow  $i$  from the network;
}
admit all remaining flows to the network for service;

```

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

Figure 4.3B Procedures of the Propose Admission Control Algorithm for New Requests

4.5 Performance Evaluation

We perform a series of simulation experiments to evaluate the performance of the DQRA+AP/CDMA protocol with the proposed scheduling algorithm for differentiated service. It shows that the proposed inter-class scheduling can effectively support the differentiated service among classes, and the particular intra-class scheduler can improve the performance of the specific class of traffic according to the corresponding requirements.

4.5.1 Experiment for Scheduling Algorithms

4.5.1.1 Experiment Design

In order to validate the operations and evaluate the performance of the proposed protocol and scheduling algorithms, discrete event simulation experiments have been conducted. In the experiments, a perfect minislot state detection is assumed so that we can concentrate on the silent feature on the scheduling issues and algorithms. The real-time messages have laxities as their time constraints. The entire network is assumed to have $N = 100$ mobiles and $K = 16$ CDMA spreading codes as logical transmission channels. The messages are generated as a Poisson process. The length of the messages follows exponential distribution with the average message length as 6400 bits. The laxity of the real-time messages also follows exponential distribution with average value of three times of the transmission time of the average message length. The time slot for data transmission is the transmission time of 640 bits. And there $m = 3$ access control minislots for each data transmission slot on each logical channel. It is different from the simulation experiments design in the pervious chapter because there is no message dropping here. Then the system cannot keep stable at the arrival rate of 16 packets/slot

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

where the network utilization is equal to one. The performance of the protocols and scheduling algorithms will be measured by the average message tardy rate defined in (4-9) and the average message delay.

4.5.1.2 Simulation Results on the Inter-class Scheduling

This experiment is established to examine the effect of the PAD scheduling algorithm as the inter-class scheduling algorithm on the integrated traffic in CDMA wireless network.

We fix the average delay ratio between the non-real-time and real-time classes to be two,

or $\frac{\bar{d}_{NRT}}{\bar{d}_{RT}} = \frac{\delta_{NRT}}{\delta_{RT}} = \frac{2}{1}$. The PAD scheduling is introduced to the DQRA+AP/CDMA

protocol in which the request collisions are resolved as Chapter 3.3.1. But in real-time and non-real-time classes the data transmission scheduling algorithms are both FIFO.

Figure 4.4 Average Message Delay vs. Total Message Arrival Rate When Real-time Traffic Percentage = 20%

Figure 4.4-4.6 show the real-time and non-real-time message delays vary as the total arrival rate when the real-time traffic percentages are 20%, 50%, and 80% respectively.

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

Figure 4.7 is a 3D graph to show the relationship among the message delay, the total arrival rate, and the real-time traffic percentage.

Figure 4.5 Average Message Delay vs. Total Message Arrival Rate When Real-time Traffic Percentage = 50%

Figure 4.6 Average Message Delay vs. Total Message Arrival Rate When Real-time Traffic Percentage= 80%

Figure 4.7 Real-time and Non-real-time Message Delay vs. Real-time Traffic Percentage and Aggregate Arrival Rate

It is clear that the inter-class scheduling is only effective in the heavy traffic. In the figures, the average delay ratio between real-time and non-real-time messages is not the expected value until the aggregate arrival rate is 15 packets/slot. When the total arrival rate is under 13 packets/slot, the ratio between the real-time and non-real-time messages is nearly equal. Table 4.1 shows the average delays of real-time and non-real-time messages in different real-time traffic percentages when the total arrival rate is 15 packets/slot.

Real-time Traffic Percentage (%)	Real-time Delay (Slots)	Non-real-time Delay (Slots)	Average Delay (Slots)
10	28.866705	49.788303	47.697056
20	26.748466	53.183292	47.898804
30	28.299707	56.478371	48.004299
40	30.037340	59.984074	48.000259
50	31.923933	63.812550	47.859585
60	34.180359	68.343292	47.829834
70	36.777706	73.533089	47.809341
80	39.765793	79.515411	47.704784
90	43.633514	87.532112	48.024540

Table 4.1 Average Messages Delay When the Aggregate Arrival Rate =15 Packets/Slot

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

In this table, the average delay is the average delay of all transmitted messages no matter what kind of traffic they belong to. This value is invariant no matter what the real-time traffic percentage is. This situation can be expected by the conservation law as equation (4-3), where the weighted average delay is only dependent on the aggregate traffic arrival rate under arbitrary distributions for the message inter-arrivals and message sizes.

On the other hand, although the ratio between the average delays fixes to be the expected value, the average delays of the real-time and non-real-time messages both increase as the real-time traffic percentage. This can also be explained by the equation (4-3). We assume that the real-time traffic percentage is α . Thus the real-time message arrival rate is $\lambda_{RT} = \alpha\lambda$, where λ is the aggregate message arrival rate to the network. And the non-real-time message arrival rate is $\lambda_{NRT} = (1-\alpha)\lambda$. Substitute λ_{RT} and λ_{NRT} to (4-3) and eliminate $\bar{L}$ in the both side because all message sizes hold the same distribution with the same average value, we have $\alpha\bar{d}_{RT} + (1-\alpha)\bar{d}_{NRT} = \bar{d}$. Here $\bar{d}$ is the average delay of all messages, which only depends on the aggregate arrival rate λ . Then we substitute

$$\frac{\bar{d}_{RT}}{\bar{d}_{NRT}} = \frac{\delta_{RT}}{\delta_{NRT}} \text{ to the previous equation, } \bar{d}_{RT} = \frac{\delta_{RT}}{\delta_{NRT} + (\delta_{RT} - \delta_{NRT})\alpha} \times \bar{d}. \text{ Since } \delta_{RT} < \delta_{NRT},$$

$\bar{d}_{RT}$ will increase as α . And $\bar{d}_{NRT}$ will also increase as α .

The authors in [24] point out that the necessary and sufficient condition for the feasibility of the PDD model when $N = 2$ is $\delta_2 \geq \frac{\bar{d}_2^{SP}}{\bar{d}_1^{SP}}$ or $\frac{1}{\delta_2} \leq \frac{\bar{d}_1^{SP}}{\bar{d}_2^{SP}}$, where $\bar{d}_i^{SP}$ is the average delay of class i in the Strict Priority (SP) scheduling, which serves class m with higher priority than class n for $m > n$. Therefore, a given delay ratio $\delta_2 < 1$ between Class 2 and Class 1

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

is feasible if and only if the corresponding delay ratio in the average delay ratio in the SP scheduling is not larger than δ_2 . Note that the average delays of the two classes in SP are a function of the aggregate arrival rate, the traffic percentage, and the aggregate average backlog, so the feasibility of the PDD model depends on the same factors.

For some queuing models, the average delays with the SP scheduling can be analytically computed. For the M/G/1 queuing model with $N = 2$ classes, for instance, $\frac{\bar{d}_1^{SP}}{\bar{d}_2^{SP}} = \frac{1}{1-\lambda}$.

This result, which is based on Poisson arrivals, shows that the feasibility range increases as the aggregate arrival rate increases. The whole system in our simulation can be modeled as an M/M/m system. The authors point out in [25] that statistical multiplexing with m channels produces a delay almost same as the delay of statistical multiplexing with the m channels combined in one for a heavy load. When the aggregate arrival rate in the simulation is 15 packets/slot, the utilization ρ is close to 1. Therefore, the delay ratio between the two classes of traffic in our system can be approximated by the delay ratio in an M/M/1 system when the inter-class scheduling is SP scheduling. Thus, we have the feasibility condition for our PAD scheduling in the simulation is

$$\frac{\delta_{NRT}}{\delta_{RT}} \leq \frac{1}{1-\lambda} = \frac{1}{1-15/16} = 16.$$

This implies that the maximum delay ratio between non-real-time and real-time messages cannot exceed 16 when the aggregated rate is 15 packets/slot.

Figure 4.8 Average Message Delay vs. Total Message Arrival Rate When Delay Ratio =4

Figure 4.9 Average Message Delay vs. Total Message Arrival Rate When Delay Ratio = 8

Figure 4.8 and 4.9 show the average message delay when the real-time traffic percentage is 50%. The delay ratios $\frac{\delta_{NRT}}{\delta_{RT}}$ are 4 and 8, respectively. It is clear that the ratios are the expected values when the aggregate arrival rate is 15 packets/slot in both situations.

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

However, the simple description on the feasibility condition as $\frac{\delta_{NRT}}{\delta_{RT}} \leq \frac{1}{1-\lambda}$ hides the dependency of the feasibility range on the class traffic percentage.

Figure 4.10 Average Message Delay vs. Total Message Arrival Rate When Real-time Traffic Percentage = 10%

Figure 4.11 Average Message Delay vs. Total Message Arrival Rate When Real-time Traffic Percentage = 30%

Figure 4.12 Average Message Delay vs. Total Message Arrival Rate When Real-time Traffic Percentage = 50%

Figure 4.13 Average Message Delay vs. Total Message Arrival Rate When Real-time Traffic Percentage = 70%

Figure 4.14 Average Message Delay vs. Total Message Arrival Rate When Real-time Traffic Percentage = 90%

When the delay ratio $\frac{\delta_{NRT}}{\delta_{RT}}$ is set to 16, or the maximum value for the feasibility condition, the simulation results in Figure 4.10-4.14 show the message delays when the real-time traffic percentages are 10%, 30%, 50%, 70%, and 90%, respectively. The results imply that the expected service differentiation can only supported when the real-time traffic percentage is high. We also establish a table as Table 4.1 to compare the average delays of real-time and non-real-time messages when the aggregate arrival rate is 15 packets/slot. The real-time traffic percentage varies from 10% to 90%.

Real-time Traffic Percentage (%)	Real-time Delay (Slots)	Non-real-time Delay (Slots)	Average Delay (Slots)
10	11.365426	51.788498	47.742676
20	11.429376	56.835560	47.763527
30	11.514565	63.590050	47.975277
40	11.674122	71.767021	47.706203
50	11.882193	83.770180	47.813297
60	12.173921	100.848145	47.593800

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

70	12.668611	129.365829	47.688248
80	13.604947	182.610291	47.434021
90	19.348764	302.006226	47.674622

Table 4.2 Average Messages Delay When the Aggregate Arrival Rate =15 Packets/Slot

From Table 4.2 we can find that the average ratio between two classes of traffic increases as the real-time traffic percentage. These results highlight the dependency of the feasibility range on the traffic percentage. The expected ratio $\frac{\delta_{NRT}}{\delta_{RT}}=16$ can only be reached when the real-time traffic percentage is 90%. Therefore, we can conclude the two requirements to obtain the maximum feasible delay ratio when there exist two classes of traffic. Firstly, the aggregate arrival rate should be close to the service capacity, or the utilization is close to 1. Secondly, the traffic percentage of higher priority class in the total traffic load should close to 100%, or the higher priority class traffic must be dominant in the whole traffic load. The maximum feasible delay ratio can be reached when these two conditions both exist.

4.5.1.3 Simulation Results on Intra-class Scheduling

In this section, the traffic in the network is a unique kind. The traffic is either all real-time traffic or all non-real-time traffic. The request collision resolution method is still the combination of fix assignment and random assignment like the DQRA+AP/CDMA protocol. However, the data transmission is decided according to the different requirements of the service.

4.5.1.3.1 Intra-class Scheduling for Real-time Traffic

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

Figure 4.15 and 4.16 show the comparison results of intra-class scheduling for real-time traffic with three disciplines, FIFO, MUF+FIFO, and the proposed one combined of MUF and SPTF. In the simulation experiments, the messages in the network are all real-time traffic with time constraints.

Figure 4.15 Real-time Message Average Tardy Rate vs. Real-time Message Arrival Rate

Figure 4.16 Real-time Message Average Delay vs. Real-time Message Arrival Rate

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

Figure 4.15 is the comparisons of message tardy rate of three scheduling schemes. When the arrival rate is 15 packets/slot, the tardy rate in FIFO (42.9%) is higher than that in MUF+SPTF (36.3%), while the tardy rate in MUF+FIFO (31.6%) is the smallest among the three. MUF outperforms FIFO in tardy rate because it schedules the message with the most urgency to service in the highest priority to lower the chance that the messages miss the deadline before the transmission. Additionally, SPTF is to improve the average message delay by allocating the message with shorter length a higher priority than a longer one. This effect on the delay performance is illustrated in Figure 4.16, where the average delay in MUF+SPTF is shorter than those in other two. Obviously, a shorter delay means a higher possibility for the message to start transmission before deadline. Thus the average tardy rate decreases as the average delay decreases. On the other hand, the average delay in MUF+FIFO is nearly same as that in FIFO, which denotes MUF discipline can only reduce the tardy rate but it has no effect on improving the delay performance.

4.5.1.3.2 Intra-class Scheduling for Non-real-time Traffic

Figure 4.17 Non-real-time Message Average Delay vs. Non-Real-time Message Arrival Rate

Then we compare the average message delays of non-real-time traffic with the intra-class schedulers use FIFO and the proposed Gated Smallest Estimated Bandwidth First (GSEBF) respectively. The simulation results are plotted in Figure 4.17. When the traffic load is very heavy with the arrival rate to the network being 15 packets/slot, the average delay in GSEBF is nearly three fifths of that in FIFO.

4.5.1.4 The Proposed Hierarchical Scheduling

As described previously, the inter-class scheduler and the intra-class schedulers work in parallel. The inter-class scheduling algorithm is the PAD algorithm. In the real-time traffic class, the messages are scheduled in MUF+SPTF order, while the messages in non-real-time class are scheduled in the GSEBF discipline. The delay ratio between two

classes of traffic is set to $\frac{\delta_{NRT}}{\delta_{RT}} = 2$.

Figure 4.18 Average Message Delay vs. Aggregate Arrival Rate When Real-time Traffic Percentage = 20%

Figures 4.18-4.20 depict the variation of average message delay as the aggregate arrival rate changes in the different real-time traffic percentages when the data transmission is

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

scheduled with the proposed hierarchical level scheduling method. Figure 4.21 is a 3D graph shows the relationship among the message delays, the real-time traffic percentage, and aggregate arrival rate.

Figure 4.19 Average Message Delay vs. Aggregate Arrival Rate When Real-time Traffic Percentage = 50%

Figure 4.20 Average Message Delay vs. Aggregate Arrival Rate When Real-time Traffic Percentage = 80%

Figure 4.21 Real-time and Non-real-time Message Delay vs. Real-time Traffic Percentage and Aggregate Arrival Rate

Similar as the simulation results in 4.5.2 where the inter-class scheduling is PAD and the intra-class scheduling for both classes of traffic is FIFO, the system can only provide expected delay differentiated service between two classes under heavy traffic. When the aggregate arrival rate is less than 13 packets/slot, the average delays of two classes are nearly equal. The real-time and non-real-time messages delays at 15 packets/slot are collected to establish Table 4.3. This table also includes the average delays of all messages in different real-time traffic percentages at 15 packets/slot.

Real-time Traffic Percentage (%)	Real-time Delay (Slots)	Non-real-time Delay (Slots)	Average Delay (Slots)
10	23.690199	33.523483	32.538200
20	18.801441	35.849834	32.440666
30	19.477478	38.159855	32.563549
40	20.695677	41.191830	32.991085
50	22.376146	44.660629	33.522945
60	24.742920	49.430504	34.621181
70	29.931496	59.818710	38.906578
80	35.270023	70.519318	42.329556

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

90	39.032780	78.298958	42.956131
----	-----------	-----------	-----------

Table 4.3 Average Messages Delay When the Aggregate Arrival Rate =15 Packets/Slot

Compare Table 4.3 and Table 4.1, it is clear that the inter-class scheduling can effectively provide the delay differentiated service when the PDD model feasibility condition exists. Both tables show that the average delay ratio between real-time and non-real-time traffic is less than the expected values when the real-time traffic percentage is smaller than 20%, where the feasibility condition is against. However, the delay ratio keeps stable as expected no matter how much the real-time traffic percentage is when the percentage is larger than 30%, where the feasibility condition can be met. Table 4.3 also implies that the inter-class scheduling will be not affected no matter what kinds of intra-class scheduling algorithms are used. But the intra-class scheduling algorithms can reduce the average message delays. Unlike Table 4.1, the average delays of all messages in Table 4.3 increase from about 32 slots to 43 slots as the real-time traffic percentages increase from 10% to 90%. However, the relationship among the three kinds of average delays always matches the equation (4-3). When the scheduling algorithm is decided, the equation shows that the average delay of all messages is only decided by the aggregate arrival rate no matter what the average delay ratio and the real-time traffic percentage are. When the real-time traffic percentage is 10%, therefore, the average delay is close to the average delay at 15 packets/slot in Figure 4.17, in which the scheduling algorithm is GSEBF. At this time, the no-real-time traffic is dominant in the total traffic. So the performance of the proposed hierarchical level scheduling is close to the system in which the messages are all non-real-time and scheduled in GSEBF order. As the real-time traffic percentage increases, the weight of the real-time traffic increases. When the percentage is close to 100%, the real-time traffic is dominant in the total traffic. Then the performance of the proposed scheduling algorithm is close to the system in which the messages are all real-

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

time and scheduled in MUF+SPTF. Note that the non-real-time traffic still affects the performance of whole system although the real-time traffic dominates the whole performance. Therefore, when the real-time traffic percentage is 90%, the delays in Table 4.3 are still smaller than in Table 4.1.

For real-time messages, the shorter average delay implies higher likelihood of successful transmission that leads to smaller tardy rate. Figures 4.22-4.24 compare the real-time message tardy rates of the original PAD scheduling with the intra-class scheduling algorithms for two traffic classes are both FIFO (RT FIFO/NRT FIFO/PAD) and the proposed hierarchical scheduling with different intra-class scheduling algorithms for different traffic classes (RT SPTF+MUF/NRT GSEBF/PAD) in different real-time traffic percentages. Figure 4.25 is a 3D graph depicts the relationship among the tardy rate, the real-time traffic percentage, and aggregate arrival rate. The results highlight that the proposed scheduling algorithm will decrease the tardy rate effectively.

Figure 4.22 Real-time Message Tardy Rate vs. Aggregate Arrival Rate When Real-time Traffic Percentage = 20%

Figure 4.23 Real-time Message Tardy Rate vs. Aggregate Arrival Rate When Real-time Traffic Percentage = 50%

Figure 4.24 Real-time Message Tardy Rate vs. Aggregate Arrival Rate When Real-time Traffic Percentage= 80%

Figure 4.25 Real-time Tardy Rate vs. Real-time Traffic Percentage and Aggregate Arrival Rate

It is clear that both the tardy rates in RT FIFO/NRT FIFO/PAD and RT MUF+SPTF/NRT GSEBF/PAD will increase as the real-time traffic percentage increases. Large real-time percentage represents more real-time messages require service at same time. Thus, the large real-time traffic percentage raises the tardy rate because of the limited service capacity. Furthermore, the difference between the tardy rates in the two different scheduling algorithms does not always increase as the real-time traffic percentage. The difference increases as the percentage when it is smaller than 60%. However, the difference decreases as the percentage increases from 60%. When the percentage is large, the real-time traffic is major in the whole traffic. But the intra-class scheduling algorithm, MUF+SPTF, for real-time traffic cannot result in a distinct reduce in the delay from that of FIFO scheduling, which has already been shown in Figure 4.16. On the other hand, the GSEBF scheduling for non-real-time traffic still lowers the average delay. This improvement still lowers the tardy rate, but its affect is diminished as the real-time percentage increases.

4.5.2 Experiment for Proposed Admission Control

For the proposed admission control based on the SEBFD discipline for real-time traffic and the SAFD discipline for non-real-time traffic, the most important procedure is to obtain the function of $\bar{D}_{RT}$ as (4-19). We use the interpolation method to estimate $\bar{D}_{RT}$ necessary for the test process in the admission control algorithm. We obtain the interpolated data based on the simulation results from the simulation in 4.5.4. Only the results in the heavy traffic are held because the proportional delay differentiator works effectively at that time.

4.5.2.1 Experiment Design

Figure 4.26 Real-time Message Delay vs. Real-time Traffic Percentage and Aggregate Arrival Rate

Figure 4.26 is a 3-dimension figure that depicts the relationship between the real-time message delay and the real-time traffic percentage at heavy aggregate arrival rate with the proposed hierarchical RT MUF+SPTF/NRT GSEBF/PAD scheduling algorithm. The

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

total arrival rate is the sum of all messages including real-time and non-real-time in the network, which increase from 14.5 packets/slot to 15.3 packets/slot. When the aggregate arrival rate is more than 15.3 packets/slot, the real-time delay is longer than 80 slots.

The admission control algorithm requires estimating $\bar{D}_{RT}$ when the users ask for service by sending requests to the base station. However, the base station or central controller only keeps the known data from the scheduling experiments. Interpolating the available data provides a solution to estimation. Interpolation is a process for estimating values that lie between known data points. It is important for applications such as signal and image processing and data visualization. The mathematic software MATLAB provides the required interpolation function, *interp2*, to the experiments for admission control. The function *interp2* performs two-dimensional interpolation. $ZI = \text{interp2}(X, Y, Z, XI, YI)$ returns matrix *ZI* containing elements corresponding to the elements of *XI* and *YI* and determined by interpolation within the two-dimensional function specified by matrices *X*, *Y*, and *Z*. *X* and *Y* must be monotonic, and have the same format ("plaid") as if they were produced by *meshgrid*. Matrices *X* and *Y* specify the points at which the data *Z* is given. Out of range values are returned as *NaNs*. Interpolation is the same operation as table lookup. Described in table lookup terms, the table is $\text{tab}=[NaN, Y; X, Z]$ and *interp2* looks up the elements of *XI* in *X*, *YI* in *Y*, and, based upon their location, returns values *ZI* interpolated within the elements of *Z*. In other words, the process of estimating $\bar{D}_{RT}$ is to look up the table composed of the know data including arrival rate and real-time traffic percentage from the scheduling experiments, and return the interpolated real-time message delay.

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

Interpolation requires that X and Y be monotonic, that is, either always increasing or always decreasing from point to point. We can prepare these matrices using the *meshgrid* function, or else be sure that the "pattern" of the points emulates the output of *meshgrid*. In addition, the interpolation automatically maps the input to an equally spaced domain before interpolating. $[X,Y] = \text{meshgrid}(x,y)$ transforms the domain specified by vectors x and y into arrays X and Y , which can be used to evaluate functions of two variables and three-dimensional mesh/surface plots. The rows of the output array X are copies of the vector x ; columns of the output array Y are copies of the vector y . Let X and Y represent the real-time traffic percentage and the arrival rate respectively.

$[X,Y]=\text{meshgrid}(0.1:0.1:0.9,14.5:0.1:15.3)$ will produce two 9×9 matrixes.

$X =$

Columns 1 through 9

0.1000	0.2000	0.3000	0.4000	0.5000	0.6000	0.7000	0.8000	0.9000
0.1000	0.2000	0.3000	0.4000	0.5000	0.6000	0.7000	0.8000	0.9000
0.1000	0.2000	0.3000	0.4000	0.5000	0.6000	0.7000	0.8000	0.9000
0.1000	0.2000	0.3000	0.4000	0.5000	0.6000	0.7000	0.8000	0.9000
0.1000	0.2000	0.3000	0.4000	0.5000	0.6000	0.7000	0.8000	0.9000
0.1000	0.2000	0.3000	0.4000	0.5000	0.6000	0.7000	0.8000	0.9000
0.1000	0.2000	0.3000	0.4000	0.5000	0.6000	0.7000	0.8000	0.9000
0.1000	0.2000	0.3000	0.4000	0.5000	0.6000	0.7000	0.8000	0.9000
0.1000	0.2000	0.3000	0.4000	0.5000	0.6000	0.7000	0.8000	0.9000

$Y =$

Columns 1 through 9

14.5000	14.5000	14.5000	14.5000	14.5000	14.5000	14.5000	14.5000	14.5000
14.6000	14.6000	14.6000	14.6000	14.6000	14.6000	14.6000	14.6000	14.6000
14.7000	14.7000	14.7000	14.7000	14.7000	14.7000	14.7000	14.7000	14.7000

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

14.8000	14.8000	14.8000	14.8000	14.8000	14.8000	14.8000	14.8000	14.8000
14.9000	14.9000	14.9000	14.9000	14.9000	14.9000	14.9000	14.9000	14.9000
15.0000	15.0000	15.0000	15.0000	15.0000	15.0000	15.0000	15.0000	15.0000
15.1000	15.1000	15.1000	15.1000	15.1000	15.1000	15.1000	15.1000	15.1000
15.2000	15.2000	15.2000	15.2000	15.2000	15.2000	15.2000	15.2000	15.2000
15.3000	15.3000	15.3000	15.3000	15.3000	15.3000	15.3000	15.3000	15.3000

Then Z represents the 9×9 matrix composed of the simulation results in Figure 4.26.

$\bar{D}_{RT}$ can be estimated by the interpolating function above with the determined arrival rate and real-time traffic percentage as following:

$$\bar{D}_{RT} = \text{interp2}(X, Y, Z, Per_{RT}, Tr_{total}), \quad (4-31)$$

The Per_{RT} and Tr_{total} can be calculated as the description in Part 4.4.2. After estimating $\bar{D}_{RT}$, the base station can decide which request should be admitted or rejected by the proposed admission control algorithm depicted in Figure 4.3A and 4.3B. In the implementation of the admission control in a practical system, the interpolation function *interp2* in MATLAB can be transplanted to the base station by other programming language, e.g. C language. The known data for the interpolation, the matrix Z in our case, can be stored in the base station with the admission control program. The rewritten interpolation program in C can estimate the $\bar{D}_{RT}$ by looking up the table, or the matrix Z .

We establish an experiment to examine our proposed admission control algorithm. The value of average arrival rate at each node is a uniform distribution between 0 packets/slot and 0.8 packets/slot. The messages are generated in a Poisson process. The inter-arrival time follows an exponential distribution with the corresponding average arrival rate from

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

the nodes. The delay requirement for the different real-time traffic flows of the real-time messages is a uniform distribution from 25 slots and 45 slots.

4.5.2.2 Experimental Results

Tables are established to collect the parameters of the messages.

Node	Average Arrival Rate (Packets/Slot)	Traffic Type	Average Delay requirement (Slots)
1	0.6963	RT	37
2	0.6977	RT	42
3	0.6093	RT	25
4	0.5356	RT	31
5	0.7216	RT	44
6	0.6572	RT	27
7	0.6661	RT	36
8	0.5119	RT	34
9	0.6265	RT	39
10	0.5851	RT	42
11	0.4518	RT	28
12	0.5787	RT	34
13	0.1918	RT	43
14	0.3919	NRT	
15	0.3389	NRT	
16	0.6365	NRT	
17	0.6107	NRT	
18	0.5081	NRT	
19	0.4928	NRT	
20	0.2148	NRT	
21	0.7930	NRT	
22	0.6083	NRT	
23	0.3857	NRT	
24	0.7562	NRT	
25	0.2886	NRT	
26	0.7588	NRT	
27	0.5527	NRT	

Table 4.4 Parameters for the Traffic Flows from the Mobile Nodes.

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

Then the parameters required for the delay approximation with equation (4-19) are calculated as the following. The aggregate arrival rate is the sum of all traffic flows is 14.8665 packets/slot. The total real-time arrival rate is 7.5296 packets/slot. Therefore, the real-time traffic percentage is 50.65%. As (4-31), $\bar{D}_{RT} = \text{interp2}(X, Y, Z, 0.5065, 14.8665) = 23.3782$. It is smaller than 25 packets/slot that is the minimum delay requirements for the real-time traffic. Thus, all traffic flows from the nodes can be admitted without violating the minimum QoS requirements for the real-time traffic.

Next, we consider another example based on the above one. Another two mobile nodes send the requests to the base station during the previous 27 nodes are in service in the network. Then the admission control works differently from the above case since it should decide whether the later two traffic flows be admitted without impairing the already admitted ones. In other words, the controller should reject both the new flows if the QoS requirements of the 27 traffic flows cannot be met after any one of two new flows admitted. In Table 4.5, Nodes 28-29 are the newly arrived ones for service.

Node	Average Arrival Rate (Packets/Slot)	Traffic Type	Average Delay requirement (Slots)
1	0.6963	RT	37
2	0.6977	RT	42
3	0.6093	RT	25
4	0.5356	RT	31
5	0.7216	RT	44
6	0.6572	RT	27
7	0.6661	RT	36
8	0.5119	RT	34
9	0.6265	RT	39
10	0.5851	RT	42

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

11	0.4518	RT	28
12	0.5787	RT	34
13	0.1918	RT	43
14	0.3919	NRT	
15	0.3389	NRT	
16	0.6365	NRT	
17	0.6107	NRT	
18	0.5081	NRT	
19	0.4928	NRT	
20	0.2148	NRT	
21	0.7930	NRT	
22	0.6083	NRT	
23	0.3857	NRT	
24	0.7562	NRT	
25	0.2886	NRT	
26	0.7588	NRT	
27	0.5527	NRT	
28	0.2689	RT	39
29	0.1048	NRT	

Table 4.5 Parameters for the Traffic Flows from the Mobile Nodes.

Currently the aggregate arrival rate will be 15.2402 packets/slot. The aggregate real-time arrival rate is 7.7985 packets/slot. The real-time traffic percentage is 51.17%. $\bar{D}_{RT} = \text{interp2}(X, Y, Z, 0.5117, 15.2402) = 68.1881$. It is obviously larger than the minimum delay requirements in all real-time traffic flows. Therefore, both new flows cannot be admitted at the same time.

However, the admission control algorithm described in Figure 4.3B but not 4.3A can be used here to make the decision. The pool for rejection only includes those latter traffic flows, which should not affect the already admitted ones. The non-real-time traffic flow is directly chosen to drop because of its smaller arrival rate comparing with the real-time

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

flow. After the dropping, the aggregate arrival rate is 15.1354 packets/slot. The total real-time arrival rate is 7.7895 packets/slot, which implies that the real-time traffic percentage is 51.47%. $\bar{D}_{RT} = \text{interp2}(X, Y, Z, 0.5147, 15.1354) = 48.5714$. It is larger than 25, the minimum delay requirement of all real-time traffic flows. Therefore, the new real-time traffic flow must also be rejected because its admission will disturb the already admitted service.

Additionally, we make the decision opposed to the above that the real-time traffic flow is rejected first because of its bigger arrival rate. The total arrival rate will be 14.9713 packets/slot, and the real-time arrival rate is 7.5296 packets/slot. Therefore, the real-time traffic percentage should be 50.29%.

$\bar{D}_{RT} = \text{interp2}(X, Y, Z, 0.5029, 14.9713) = 29.3692$. This estimate delay is still bigger than the minimum delay requirement of the real-time traffic flows. Finally, the new non-real-time traffic flow must be rejected to meet the QoS requirements of the traffic flows. Both new flows should both be rejected as above.

The third experiment is different from the pervious two. It can only support the QoS requirement for the real-time traffic after dropping some of the original flows when all the flows send the simultaneous requests at the beginning.

Node	Average Arrival Rate (Packets/Slot)	Traffic Type	Average Delay requirement (Slots)
1	0.6083	RT	35
2	0.6367	RT	28
3	0.2511	RT	26
4	0.1887	RT	42
5	0.3956	RT	39

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

6	0.5482	RT	45
7	0.7557	RT	41
8	0.3514	RT	40
9	0.3979	RT	36
10	0.4436	RT	43
11	0.3486	RT	27
12	0.7470	RT	29
13	0.7988	RT	42
14	0.5525	RT	39
15	0.7688	NRT	
16	0.6204	NRT	
17	0.2604	NRT	
18	0.4353	NRT	
19	0.2452	NRT	
20	0.5382	NRT	
21	0.2766	NRT	
22	0.7845	NRT	
23	0.2832	NRT	
24	0.6502	NRT	
25	0.7863	NRT	
26	0.2495	NRT	
27	0.3031	NRT	
28	0.5984	NRT	
29	0.3543	NRT	
30	0.7941	NRT	

Table 4.6 Parameters for the Traffic Flows from the Mobile Nodes.

The total arrival rate is 14.9724 packets/slot. The real-time message arrival rate is 7.0240 packets/slot. The real-time traffic percentage is 46.91%.

$\bar{D}_{RT} = \text{interp2}(X, Y, Z, 0.4691, 14.9724) = 28.9762$. It is larger than 26 slots that is the minimum delay requirement for real-time traffic.

According to the proposed admission control algorithm in Figure 4.3A, the Node 3 has the minimum estimated buffer size in all real-time nodes. But its arrival rate is 0.2511

Chapter 4 Providing Differentiated Services in CDMA Wireless Networks

packets/slot that is bigger than 0.2452 packets/slot of the Node 19, which is the minimum of all non-real-time traffic flows. Therefore, the traffic flow from Node 19 will be rejected. Then the parameters for the delay estimation are all updated. The aggregate arrival rate is 14.7272 packets/slot. The real-time traffic percentage is 47.69%. $\bar{D}_{RT} = \text{interp2}(X, Y, Z, 0.4769, 14.7272) = 18.7230$. It is less than the minimum delay requirement for real-time traffic. So the rest traffic flows will be admitted to support the real-time traffic after the Node 19 is blocked out of service.

These three experiments show the proposed admission control algorithm can support the QoS requirements for real-time traffic in the proposed hierarchical scheduling. At same time, it will control the network utilization as high as possible.

Chapter 5 Conclusions and Recommendations

5.1 Conclusions

We have presented our solutions to two issues, integrated and differentiated services for the research work of MAC protocols in CDMA wireless networks. Based on the original DQRAP/CDMA protocol that can achieve a very good performance in supporting data traffic, we design the new collision resolution and data transmission scheduling algorithms to provide QoS service for the integrated traffic in the future wireless networks.

5.1.1 Providing Integrated Service in CDMA Networks

Firstly, we have proposed a novel MAC protocol, named as distributed queuing random access with assignment protocol (DQRA+AP/CDMA), and corresponding scheduling algorithms to provide integrated service to the mixed traffic composed of real-time and non-real-time messages in a CDMA wireless network. The modification to the collision resolution scheme is to combine the random access scheme with the fixed assignment method to solve the collisions among the transmission requests when they access the control subslots for supporting real-time service. Then a time-based priority scheme has been proposed, which is named as MUF, to allocate the higher priorities to the real-time messages with the more stringent time constraints in order to reduce the average message loss rate due to deadline expiry. The MUF, however, could cause longer delays to the messages with larger laxities. We have introduced FIFO principle to integrate with MUF to make that the messages are scheduled in MUF order in a short term while in a long

Chapter 5 Conclusions and Recommendations

term, messages are transmitted referring the arrival time. Moreover, this MUF-FIFO scheme has been extended to accommodate integrated traffic composed of real-time messages with time constraints and non-real-time messages without time constraints to achieve a reasonable performance behavior of the network. Two more complicated scheduling algorithms, MUFC and CBPA, have been proposed based on MUF-FIFO to enhance the performance of the integrated service with certain fairness. MUFC is to insert at most one non-real-time message between the successive deliveries of two real-time messages. CBPA considers the message length and laxity to be the factors to the cost of message transmission. The dynamic priority of a message is defined by its normalized laxity that represents the urgency of the message, and the normalized message length that represents the message transmission time. The message with the smallest value of the priority is to be transmitted first. In the simulation experiments, the major metric to evaluate real-time service is the average message loss rate while the major metric to evaluate network service is the average message delay. The simulation results indicate that the MUFC and CBPA can better support integrated service than MUF-FIFO. Furthermore, CBPA has the best performance among the three algorithms. It reveals that the proposed scheduling algorithms can have a good performance service for real-time traffic while improving the performance of the service to non-real-time traffic.

5.1.2 Providing Differentiated Service in CDMA Networks

We have taken PAD algorithm as our inter-class scheduler that provides the delay differentiation between real-time and non-real-time traffic because it has no limitation on the queuing discipline for intra-class scheduler. The intra-class scheduler for real-time traffic is composed of MUF and SPTF queuing disciplines so that it can transmit a message with a more urgency earlier to reduce the average message tardy rate, while at

Chapter 5 Conclusions and Recommendations

the same time, it will not increase the average message delay very much. On the other hand, the intra-class scheduler for non-real-time class is taking a GSEBF order for message transmission so that the average message delay can be lowered as much as possible without the starvation problem. The proposed admission control scheme based on the combined hierarchical scheduler is to limit the number of the admitted traffic flows to the network. Thus, the time requirements of the real-time traffic flows can be met and the network utilization can also be kept in as much as possible. The simulation results indicate that the inter-class and the intra-class schedulers can work effectively in parallel. The intra-class schedulers can efficiently improve the performance of the services to real-time and non-real-time traffic according to their particular demands. The experiments on the admission control algorithm present that the experimental function to estimate the average message delay for real-time traffic is effective enough to support the admission control. The admission control algorithm can provide a guarantee to the QoS requirements of the real-time traffic by limiting the number of the accepted traffic flows.

5.2 Recommendations

Several new solutions have been introduced and developed in this thesis. However, there is still an ample scope to extend the current research work. A few specific topics are suggested below.

5.2.1 Integrated Service

- Newly designed methods are needed to support the integrated service for a more complex traffic environment including multimedia traffic such as MPEG traffic. The physical layer factor such as channel error rate and near-far problem should

Chapter 5 Conclusions and Recommendations

be considered when designing the novel protocols so that the research work can match the practical condition in a greater degree.

- An admission control algorithm is necessary to provide guaranteed service to the integrated traffic. And a mathematical model for the scheduling discipline is needed to accurately support the admission control algorithm.

5.2.2 Differentiated Service

- Extend the network traffic from two coarse classifications to a more reasonable condition that can represent the multimedia traffic with different QoS requirements in practice. Design more delicate intra-class schedulers to better support the particular requirements of the different classes.
- Develop a mathematical model to evaluate the performance of the proposed protocol. It is a prerequisite for the admission control algorithm work accurately, which is expected to provide a more precise guarantee to the system than the experimental estimation used currently.

Author's Publication

Published

Maode Ma, and Qichao. Zhu, "Providing Real-Time Service in CDMA Wireless Networks," *Proceedings of the 4th International Workshop on ITS Telecommunications*, Jul. 2004, pp. 309-313.

Maode Ma, and Qichao Zhu, "Providing Integrated Service in CDMA Wireless Networks," *Proceedings of the Fifth International Conference on Information, Communications and Signal Processing (ICICS 2005)*, Dec. 2005.

Maode Ma, and Qichao Zhu, "Providing Integrated Service in CDMA Wireless Networks," accepted by the *International Journal of Computer and Electrical Engineering*.

In Submission

Maode Ma, and Qichao Zhu, "Providing Real-Time Service in CDMA Wireless Networks," to the *International Journal of Wireless Personal Communications*.

Bibliography

- [1] I. F. Akyildiz and J. McNair *et al.*, "Medium Access Control Protocols for Multimedia Traffic in Wireless Networks," *IEEE Network*, Vol.17, Issue 4, Jul.-Aug. 2003, pp. 39-47.
- [2] R. Pichna and Q. Wang, "A Medium-Access Control Protocol for a Cellular Message CDMA Carrying Multirate Traffic," *IEEE Journals on Selected Area in Communications*, Vol. 14, No. 9, Dec. 1996, pp. 1728-1736.
- [3] C. Roobol, P. Beming, J. Lundsjo, and M. Johansson, "A Proposal for an RLC/MAC Protocol for Wideband CDMA Capable of Handling Real Time and Non Real Time Services," *Proceedings of IEEE Vehicular Technology Conference 1998*, pp. 107-111.
- [4] In-Taek Lim, "A MAC Protocol for Integrated Voice/Data Services in the Local Wireless CDMA Communication Networks," *IEEE Transactions on Consumer Electronics*, Vol. 46, No. 2, May 2000, pp. 318-325.
- [5] K. Su Park, and Dong Ho Cho, "An Advanced Channel Access Scheme for Integrated Multimedia Services with Various Bit Rates in CDMA Networks," *IEEE Communications Letters*, Vol. 3, No. 4, Apr. 1999, pp. 91-93.
- [6] B. E. Brand and A. H. Aghvami, "Multidimensional PRMA with Prioritized Bayesian Broadcast—A MAC Strategy for Multiservice Traffic over UMTS," *IEEE Transactions on Vehicular Technology*, Vol. 47, No. 4, Nov. 1998, pp.1148—1161.

- [7] J. P. Romero, R. Agustí, and O. Sallent, "An Adaptive ISMA-DS/CDMA MAC Protocol for Third-Generation Mobile Communications Systems," *Proceedings of IEEE Transactions on Vehicular Technology*, Vol. 50, No. 6, Nov. 2001, pp. 1354-1365.
- [8] L. Lenzini, M. Luise, and R. Reggiannini, "CRDA: A Collision Resolution and Dynamic Allocation MAC Protocol to Integrate Data and Voice in Wireless Networks," *IEEE Journals on Selected Area in Communications*, Vol. 19, No. 6, Jun. 2001, pp. 1153-1163.
- [9] I. F. Akyildiz, D. A. Levine, and I. Joe, "A Slotted CDMA Protocol with BER Scheduling for Wireless Multimedia Networks," *IEEE/ACM Transactions on Networking*, Vol. 7, No. 2, Apr. 1999, pp. 146-158.
- [10] V. Huang, and W. Zhuang, "QoS Oriented Message Scheduling for Wireless Multimedia CDMA Communications," *IEEE Transactions on Mobile Computing*, Vol. 3, No. 1, Jan.-Feb. 2004, pp. 73-85.
- [11] C. C. Chang, K. Ch. Chen, "Medium Access Protocol Design for Delay-Guaranteed Multicode CDMA Multimedia Networks," *IEEE Transactions on Wireless Communications*, Vol. 2, No. 6, Nov. 2003, pp. 1159-1167.
- [12] X. Wang, "Wide-Band TD-CDMA MAC With Minimum-Power Allocation and Rate- and BER-Scheduling for Wireless Multimedia Networks," *IEEE/ACM Transactions on Networking*, Vol. 12, No. 1, Feb. 2004, pp. 103-116.
- [13] D. J. Goodman, R. A. Valenzuela, K. T. Gayliard, and B. Ramamurthi, "Message Reservation Multiple Access for Local Wireless Communications," *IEEE Transactions on Communications*, Vol. 37, No. 8, Aug. 1989, pp. 885-890.

Bibliography

- [14] M. J. Karol, Z. Liu, and K. Y. Eng, "Distributed-Queuing Request Update Multiple Access (DQRUMA) for Wireless Message (ATM) Networks," *Proceedings of IEEE International Conference on Communications 1995*, Seattle, WA, pp. 1224–1231.
- [15] W. Xu and G. Campbell, "A Near Perfect Stable Random Access Protocol for a Broadcast Channel," *Proceedings of IEEE International Conference on Communications 1992*, Vol. 1, pp. 370–374.
- [16] W. Xu and G. Campbell, "DQRAP—A distributed Queuing Random Access Protocol for a Broadcast Channel," *Proceedings of Special Interest Group on Data Communication (SIGCOMM) 1993*, San Francisco, Sept. 14, 1993, pp. 270-278.
- [17] L. Alonso, R. Agustic, and O. Sallent, "A Near-Optimum MAC Protocol Based on the Distributed Queuing Random Access Protocol (DQRAP) for a CDMA Mobile Communication System," *IEEE Journals on Selected Area in Communications*, Vol. 18, No. 9, Sep. 2000, pp. 1701-1718.
- [18] G. Giovannini, L. Alonso, and R. Agusti, "Proposal of DQRAP/CDMA MAC Protocol Optimization," *Proceedings of IEEE Vehicular Technology Conference 2001*, Vol. 2, May 2001, pp. 851 –855.
- [19] S. Malarvizhi, K. Ramani, H. Rajasekaran, and M. Meenakshi, "A Near Optimum MAC Protocol Based on the Incremental Collision Resolution Multiple Access Protocol for CDMA Based Communications System," *Proceedings of 5th International Symposium on Wireless Personal Multimedia Communications*, Vol. 3, Oct. 2002, pp. 902 – 906.
- [20] M. D. Ma, and Q. C. Zhu, "Providing Real-Time Service in CDMA Wireless Networks," *Proceedings of the 4th International Workshop on ITS Telecommunications*, Jul. 2004, pp. 309-313.

- [21] M. D. Ma, and X. H. Huang, "Adaptive Scheduling for Integrated Traffic on WDM Optical Networks," *Computer Networks*, Vol. 44, No. 6, 2004, pp. 773 - 788.
- [22] O. Rose, "The Shortest Processing Time First (SPTF) Dispatch Rule and Some Variants in Semiconductor Manufacturing," *Proceedings of Simulation Conference '01*, Vol. 2, Dec. 2001, pp. 1220 - 1224.
- [23] M. Moser, and S. Engell, "A Survey of Priority Rules for FMS Scheduling and their Performance for the Benchmark Problem," *Proceedings of the 31st IEEE Conference on Decision and Control*, Vol. 1, Dec. 1992, pp. 392 - 397.
- [24] C. Dovrolis, D. Stiliadis, and P. Ramanathan, "Proportional Differentiated Services: Delay Differentiation and Packet Scheduling," *IEEE/ACM Transactions on Networking*, Vol. 10, No. 1, Feb. 2002, pp. 12-26.
- [25] C. Dovrolis, and P. Ramanathan, "A Case for Relative Differentiated Service and the Proportional Differentiation Model," *IEEE Network*, Vol. 13, Sep.-Oct. 1999, pp. 26-34.
- [26] L. Kleinrock, *Queuing systems*. New York: Wiley, 1976, Vol. II.
- [27] G. Bolch, S. Greiner, H. Meer, and K. S. Trivedi, *Queuing Networks and Markov Chains*. New York: Wiley, 1999.
- [28] D. Bertsekas, and R. Gallager, *Data Networks*. New Jersey: Prentice Hall, 1987.
- [29] C. Dovrolis, and P. Ramanathan, "Dynamic Class Selection: From Relative Differentiation to Absolute QoS," *Proceedings of International Conference on Network Protocols*, Nov. 2001, pp. 120-128.
- [30] Y. E. Liu, and J. W. Wang, "Admission Control in Deadline-Based Network Resource Management," *Proceedings of IEEE International Conference on Performance, Computing, and Communications 2004 (IPCCC '04)*, April 2004, pp.95-102.

Bibliography

- [31] R. Rietman, and J. Resing, "An $M/G/1$ Queueing Model with Gated Random Order of Service," *Queueing Systems*, Vol. 48, 2004, pp. 89-102.
- [32] T. Janevski, *Traffic Analysis and Design of Wireless IP Networks*. Massachusetts: Artech House, 2003.

APPENDIX

A Brief Introduction to SSS

The Simulation Subroutine Set (SSS) library is a software tool to help building discrete simulation programs by C language. It is originated by Dr. M. A. Pollatschek in Israel that is a professor of operation research and a consult on using computer simulation for planning and decision support in manufacturing, service, and commercial enterprises.

Followings are the description of some functional grouping of routines in SSS used for our simulation experiments in the research work:

$A(n)$	Returns the n th attribute of the current entity
$AIQ(m, r, n)$	Returns the n th attribute of the r th entity in queue m
$CREATE(g, i)$	Creates an arrival in time g from present time with identity i
$DISPOS()$	Destroys the current entity if one exists
$EX(M)$	Returns a value depending on the exponential distribution with mean M
$IDE()$	Returns the identity of the current entity
$INIQUE(q, b, s)$	Initializes q queues, b attributes, and s statistics
$NEXTEV()$	Returns the event code of the following event
$NQ(m)$	Returns the number of entities in queue m
$QUEUE(m, p)$	Enters the current entity to queue m with priority p

Appendix

$RA()$	Returns a random value depending on the uniform distribution between 0 and 1
$REMOVQ(r)$	Removes the r th entity from queue m
$SCHED(g, e, i)$	Schedules event $e > 1$ for time g from present time with identity i
$SETA(n, v)$	Sets the n th attribute's value of the current entity to v
$SETIDE(i)$	Sets the identity of the current entity to i
$SETQDC(m, d)$	Sets discipline d in the queue m ; d is either <i>FIFO</i> , <i>LIFO</i> , <i>SVF</i> , or <i>BVF</i>
$SIMEND(g)$	Ends simulation at time g or immediately if $g = 0$
$T()$	Returns the current simulated time
$UN(I, C)$	Returns a value depending on the uniform distribution between I and C